



Institution-Level Readiness Modeling for Auditable Clinical LLM Adoption in US Health Systems

Laura Kunz

Graduate School for Cellular and Biomedical Sciences, University of Bern, Bern, Switzerland

Abstract

The rapid advancement of artificial intelligence, particularly large language models, presents a transformative opportunity for the healthcare sector. However, the deployment of these complex generative systems within clinical environments introduces unprecedented challenges regarding patient safety, regulatory compliance, data privacy, and institutional accountability. This paper proposes a comprehensive institution-level readiness modeling framework designed specifically for the auditable adoption of clinical large language models within United States health systems. By synthesizing principles from organizational theory, health informatics, and regulatory science, the proposed framework delineates the multidimensional prerequisites necessary for safe integration. The study explores the technical infrastructural demands, data governance protocols, workforce competency requirements, and the fundamental necessity for transparent auditing mechanisms capable of continuously monitoring model behavior. Through simulated application across diverse healthcare settings, the research highlights critical gaps in current institutional preparedness and provides strategic recommendations for addressing these vulnerabilities. The findings underscore that successful integration relies not merely on the technical sophistication of the artificial intelligence models but fundamentally on the robust, auditable operational architecture established by the adopting institution. Ultimately, this framework provides healthcare administrators, clinical leaders, and policymakers with a systematic approach to navigate the complexities of generative artificial intelligence adoption while safeguarding clinical integrity.

Keywords: *Clinical Large Language Models, Institutional Readiness, Health Systems Auditing, Adoption Frameworks*

1. Introduction

1.1 Background and Context

The contemporary landscape of the United States healthcare system is characterized by an overwhelming accumulation of unstructured clinical data, an escalating demand for administrative efficiency, and a persistent necessity to elevate the quality of patient care. In this demanding environment, the emergence of advanced generative artificial intelligence, most notably large language models, has precipitated a paradigm shift in how health systems conceptualize clinical documentation, diagnostic support, and patient communication. Early exploratory applications of these models have demonstrated remarkable capabilities in summarizing complex medical histories, drafting routine clinical correspondence, and even passing standardized medical licensing examinations [1]. The underlying architecture of these models, predicated on deep neural networks trained on vast corpora of diverse textual data, allows them to process and generate human-like text with an unprecedented degree of contextual nuance and semantic accuracy [2]. Consequently, healthcare technology vendors and institutional innovation departments are aggressively pursuing the integration of these models into electronic health records and daily clinical workflows. Despite the immense potential, the transition

from experimental success to ubiquitous clinical deployment remains heavily constrained by the structural and regulatory complexities inherent to health systems [3]. The United States healthcare sector is governed by stringent frameworks designed to protect patient privacy and ensure clinical efficacy, primarily operationalized through the Health Insurance Portability and Accountability Act and the regulatory oversight of the Food and Drug Administration. Generative artificial intelligence systems, by virtue of their probabilistic nature and opacity, introduce substantial risks that conflict with these established paradigms [4]. Models are known to produce factually inaccurate statements, commonly referred to as hallucinations, which in a clinical context could lead to severe adverse patient outcomes. Furthermore, the capacity for these systems to inadvertently memorize and reproduce protected health information from their training datasets presents a profound liability [5]. Therefore, the enthusiastic pursuit of large language model integration must be balanced with a rigorous, systematic evaluation of institutional capacity to manage, monitor, and mitigate these emerging risks.

1.2 Problem Statement and Objectives

The central problem addressed in this research is the critical absence of standardized, comprehensive frameworks capable of assessing and guiding the institutional readiness of health systems for the auditable deployment of clinical large language models. While significant academic and industry effort has been directed toward improving the technical performance and safety guardrails of the models themselves, there has been a disproportionate lack of focus on the operational, infrastructural, and governance readiness of the organizations that intend to use them. An institution may procure a state-of-the-art, medically fine-tuned language model, but without the appropriate data pipelines, monitoring software, clinical governance committees, and workforce training protocols, the deployment will inevitably fail to meet the rigorous standards of clinical auditing [6]. Auditability in this context is defined as the continuous, transparent, and reproducible evaluation of the model inputs, outputs, operational performance, and clinical impact over time. Without comprehensive auditability, health systems cannot defend their utilization of artificial intelligence in the event of medical malpractice litigation or regulatory investigation. This paper seeks to address this critical gap by developing and detailing an Institution-Level Readiness Model specifically tailored for auditable large language model adoption. The primary objective is to articulate a multidimensional framework that deconstructs the concept of readiness into actionable, measurable domains. This involves identifying the specific technological infrastructure required to securely host and interface with these models, the data governance policies necessary to manage prompts and responses containing protected health information, and the clinical workflow redesigns essential for human-in-the-loop oversight [7]. Additionally, the research aims to provide a robust methodology for continuous institutional auditing, ensuring that model performance remains aligned with clinical standards as organizational practices evolve and model drift occurs. By establishing this foundational model, the paper provides a crucial strategic tool for healthcare executives, chief medical information officers, and regulatory bodies seeking to navigate the safe and responsible institutionalization of generative artificial intelligence.

2. Literature Review and Conceptual Background

2.1 Evolution of Artificial Intelligence in Healthcare

To adequately contextualize the institutional challenges associated with deploying large language models, it is essential to trace the historical trajectory of artificial intelligence integration within clinical settings. The initial wave of clinical artificial intelligence was dominated by rule-based expert systems in the late twentieth century, which required extensive manual encoding of medical knowledge by domain experts [8]. While these systems provided foundational insights into computational diagnostic support, they were fundamentally limited by their inability to scale, their fragility when encountering edge cases, and their failure to integrate seamlessly into clinical workflows. These early failures highlighted a critical lesson for institutional readiness: technological sophistication cannot supersede workflow integration and human-centric design. Subsequently, the digitization of clinical data through widespread electronic health record adoption created the necessary substrate for the second wave of artificial intelligence, characterized by machine learning and deep learning methodologies applied primarily to structured numerical data and medical imaging [9].

During this second wave, predictive models for sepsis detection, readmission risk scoring, and image-based diagnostic triage became increasingly prevalent. Institutions adapted to these technologies by establishing distinct operational pipelines for tabular and imaging data, typically characterized by batch processing and delayed inferencing. The governance models developed during this era were predicated on deterministic algorithms; if an algorithm classified an image in a specific manner, the reasoning could often be traced back to distinct features or statistical probabilities related to defined cohorts [10]. However, the institutional infrastructure and governance committees optimized for these specific, narrow artificial intelligence tasks are wholly inadequate for the third wave of artificial intelligence characterized by generative large language models. Unlike predictive algorithms that output a binary classification or a continuous risk score, large language models output highly variable, unstructured text that must be interpreted contextually by the clinician in real-time [11]. This shift from deterministic classification to probabilistic generation requires a complete reconceptualization of how institutions validate software, train their personnel, and monitor for errors, thereby necessitating the development of novel readiness paradigms.

2.2 Challenges in Auditable Clinical Deployments

The fundamental barrier to the widespread adoption of large language models in United States health systems is the difficulty of establishing rigorous auditability mechanisms. In traditional software deployment, auditability is achieved through deterministic testing, where a specific input is guaranteed to produce a specific output, allowing for the comprehensive mapping of system behavior before clinical release [12]. Generative models, conversely, are non-deterministic; the identical clinical prompt submitted at different times or with microscopic variations in phrasing can yield substantially different responses. This probabilistic nature severely complicates the establishment of baselines for safety and efficacy [13]. Furthermore, the deployment strategies currently favored by many institutions rely on cloud-based application programming interfaces provided by third-party vendors. This architecture creates a black-box environment where the health system has no visibility into the underlying model weights, training data updates, or hidden prompt injections executed by the vendor, rendering institutional auditing exceptionally difficult [14].

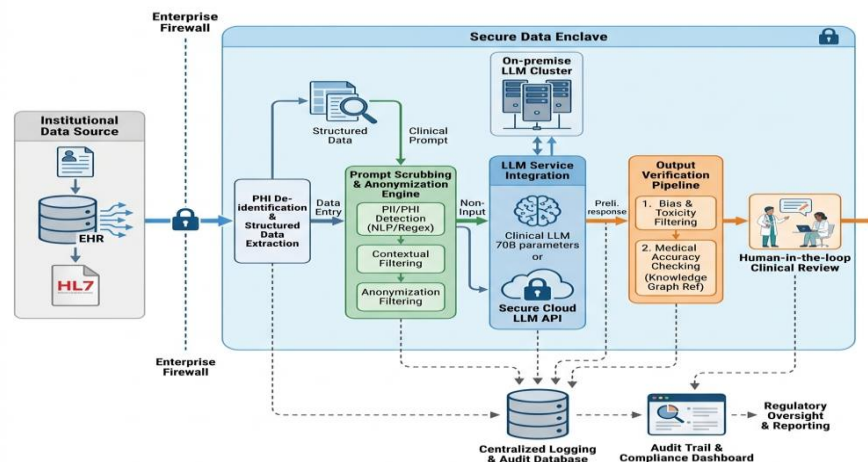


Figure 1 Comprehensive Schematic of the Institutional Data Flow and Auditing Architecture for Clinical Large Language Models

Another critical challenge to auditability is the phenomenon of semantic drift and knowledge obsolescence. Medical knowledge is continuously evolving, with new guidelines, pharmaceuticals, and protocols emerging rapidly. A large language model trained on data up to a specific cutoff date will systematically fail to provide accurate clinical insights based on subsequent discoveries. Institutions must therefore develop readiness capabilities to augment model generation with contemporary institutional data, typically through methodologies such as retrieval-augmented generation. However,

implementing retrieval-augmented generation requires sophisticated institutional data indexing, robust search infrastructure, and continuous validation of the source documents being retrieved [15]. Without these structural components, the auditing process cannot differentiate between an error caused by a hallucination within the model itself and an error caused by the retrieval of an outdated institutional policy document. Additionally, the regulatory environment surrounding these deployments remains fluid. Auditable readiness requires the establishment of dynamic governance structures capable of adapting to shifting guidance from the Office of the National Coordinator for Health Information Technology and evolving liability frameworks concerning artificial intelligence malpractice [16].

3. Methodology and Readiness Modeling Framework

3.1 Structural Components of the Readiness Model

The Institution-Level Readiness Model developed in this research is structured around a multi-pillar architecture, designed to provide a holistic assessment of a health system's capacity to deploy and audit clinical large language models safely. The methodology for constructing this model involved synthesizing best practices from high-reliability organizations, health informatics implementation science, and cybersecurity risk management frameworks. The first structural pillar is Technological and Computational Infrastructure. This domain evaluates the physical and virtual hardware capabilities required to support either on-premises model hosting or secure cloud integrations. Institutions must demonstrate sufficient graphical processing unit provisioning for low-latency inference, robust network bandwidth capable of handling the continuous streaming of text data to and from the electronic health record, and dedicated secure enclaves for processing queries containing protected health information [17]. Readiness in this pillar also mandates the existence of specialized application programming interface gateways that rate-limit requests, encrypt data in transit and at rest, and maintain immutable logs of all system interactions for post-incident forensic analysis. The second pillar focuses on Data Architecture and Governance. A health system cannot safely deploy generative models without absolute control over the data feeding into them and the data produced by them. This component models the readiness of an institution to execute real-time de-identification of clinical text before it reaches external model providers, a critical requirement for maintaining regulatory compliance. It further evaluates the maturity of the institutional knowledge bases utilized for retrieval-augmented generation workflows. An auditable deployment requires that all institutional guidelines, formulary lists, and clinical pathways are digitized, semantically indexed, and governed by a strict version control system. If a clinician uses an artificial intelligence tool to draft a discharge summary based on internal policy, the auditing framework must be able to instantly recall the exact version of the policy document that was fed to the model at the exact timestamp of the query.

Table 1 Institutional Readiness Pillars and Assessment Metrics

Readiness Pillar	Core Assessment Domain	Key Performance Indicator for Auditability
Technical Infrastructure	Hardware and Network Latency	Inference time under 500ms; 99.9% uptime log retention
Data Governance	Real-time De-identification	Zero PHI leakage events in continuous red-team testing
Clinical Workflow	Human-in-the-loop Integration	100% of generated clinical notes require physician sign-off
Regulatory Compliance	Policy and Version Control	Complete traceability of RAG context documents via hashing
Workforce Competency	AI Literacy and Prompting Skills	Mandatory annual certification in LLM hallucination detection

The third and fourth pillars are Clinical Workflow Integration and Regulatory Compliance, respectively. Workflow integration readiness necessitates a paradigm shift from passive software consumption to active, human-in-the-loop collaboration. The model evaluates whether an institution has designed its user interfaces to explicitly flag artificially generated text, preventing automation bias where clinicians blindly trust model outputs [18]. It also assesses the presence of frictionless feedback mechanisms allowing users to instantly report inaccurate or biased model responses, thereby creating a continuous human-driven auditing loop. Regulatory compliance readiness involves the establishment of multidisciplinary artificial intelligence oversight committees comprising clinicians, data scientists, legal counsel, and patient advocates. This committee is responsible for approving specific use cases, defining the parameters of acceptable risk, and conducting periodic audits of model performance across diverse patient demographics to ensure equitable care delivery.

3.2 Evaluation Metrics and Institutional Auditing

To operationalize the readiness model, a rigorous system of evaluation metrics and continuous auditing protocols must be implemented. Theoretical readiness is insufficient; health systems require quantitative measures to track their progress and identify operational vulnerabilities. The proposed methodology introduces a tiered scoring system that assesses institutional maturity across the aforementioned pillars. At the foundational tier, institutions are evaluated on compliance-based metrics, such as the implementation of basic data encryption, the completion of vendor security assessments, and the establishment of preliminary usage policies. Progression to the intermediate tier requires demonstration of active operational monitoring, including the ability to track model latency, error rates, and user adoption metrics in real-time.

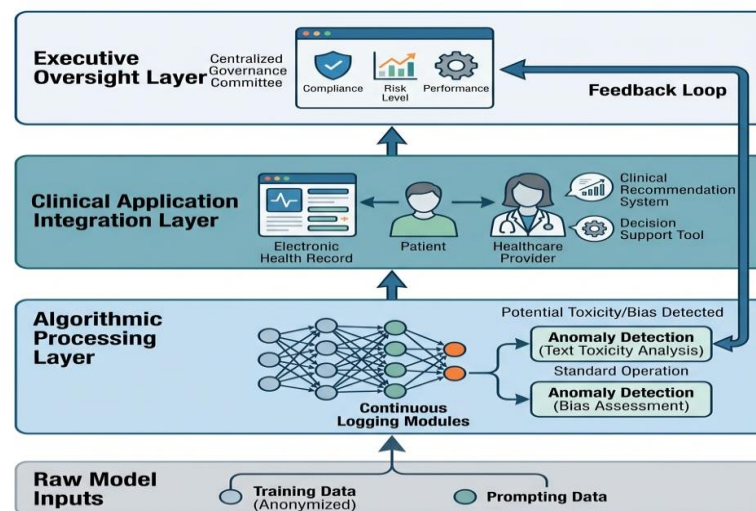


Figure 2 Multilayered Institutional Auditing Pipeline for Generative Artificial Intelligence

The advanced tier of readiness, which is strictly required for fully auditable clinical deployment, mandates the implementation of automated, continuous evaluation pipelines. Given the volume of text generated in a clinical setting, manual auditing by human reviewers is mathematically unfeasible. Therefore, the institution must deploy secondary, smaller diagnostic language models specifically trained to evaluate the outputs of the primary clinical model [19]. These evaluator models continuously scan generated clinical notes for contradictions with the patient's electronic health record, instances of severe medical hallucinations, and deviations from institutional guidelines. Any flagged anomalies are then routed to specialized clinical informatics teams for human review. The metrics derived from this continuous auditing process include the hallucination detection rate, the time-to-resolution for identified errors, and the frequency of user-initiated corrections. By tracking these metrics longitudinally, the institution can maintain a dynamically updated readiness score that accurately reflects its capacity to manage the evolving risks associated with generative artificial intelligence.

4. Results and Discussion of Framework Application

4.1 Simulated Institutional Readiness Outcomes

To validate the utility of the Institution-Level Readiness Model, a comprehensive simulation was conducted applying the framework to five distinct archetypes of United States health systems: a large academic medical center, a regional community hospital network, a specialized pediatric facility, a rural critical access hospital, and a massive multi-state corporate health system. The simulated application utilized aggregated operational data, historical technology adoption rates, and standardized compliance self-assessments to project the readiness scores of these diverse organizational structures. The results demonstrated massive disparities in preparedness, revealing that technological enthusiasm frequently outpaces infrastructural and governance realities. The large academic medical center and the multi-state corporate system exhibited the highest overall readiness scores, primarily due to their preexisting investments in extensive cloud infrastructure, established clinical informatics departments, and robust legal teams capable of negotiating complex vendor agreements.

Table 2 Simulated Readiness Scores Across Diverse Health System Archetypes

Health System Archetype	Infrastructure Score (0-100)	Governance Score (0-100)	Auditing Capacity Score (0-100)	Overall Readiness
Academic Medical Center	92	88	85	High Preparedness
Multi-State Corporate	89	91	82	High Preparedness
Pediatric Specialty	75	85	60	Moderate Vulnerability
Regional Community	60	55	40	Severe Vulnerability
Rural Critical Access	45	40	25	Critical Deficiency

Despite their high overall scores, the simulation identified significant vulnerabilities within the auditing capacity of even the most advanced systems. While the academic center excelled in hardware provisioning and model customization, its decentralized departmental structures hindered the implementation of a unified, enterprise-wide auditing pipeline. Different departments were simulating the deployment of disparate models from various vendors, creating a fragmented data landscape that made continuous cross-institutional performance evaluation nearly impossible. Conversely, the regional community hospital network and the rural critical access hospital demonstrated critical deficiencies across nearly all pillars. These organizations lacked the necessary capital to overhaul their legacy IT networks, possessed minimal in-house data science expertise, and relied heavily on generic, unmodified electronic health record vendor solutions that offered little to no transparency regarding the underlying generative algorithms. The simulation highlights that without targeted intervention and structural support, the unguided adoption of clinical large language models will significantly exacerbate the existing technological and quality-of-care divide between well-resourced institutions and vulnerable healthcare safety nets.

4.2 Strategic Implications for US Health Systems

The analysis of the readiness model and the simulated outcomes generate profound strategic implications for executive leadership within United States health systems. The most critical realization is that the deployment of generative artificial intelligence cannot be treated as a standard software acquisition. It requires a fundamental restructuring of institutional operations and an ongoing commitment to algorithmic governance. Chief Information Officers and Chief Medical Information Officers must transition from a mindset of procurement to a mindset of continuous orchestration. This involves making complex strategic decisions regarding the balance between open-source models hosted internally, which offer maximum auditability and data control but require immense technical overhead, and proprietary cloud-based models, which offer superior out-of-the-box performance but severely

limit institutional transparency. Furthermore, the requirement for robust auditability has significant financial implications. The hidden costs of large language model adoption do not reside solely in the application programming interface token charges or the software licensing fees; the vast majority of the financial burden stems from establishing the readiness infrastructure. Health systems must budget for comprehensive data cleaning initiatives to prepare institutional knowledge bases for retrieval-augmented generation, the hiring of specialized artificial intelligence assurance personnel, and the increased time required for clinicians to review and verify model outputs during the initial deployment phases. Institutions that attempt to bypass these foundational readiness investments to achieve rapid deployment will inevitably face compounded costs associated with remediation, regulatory fines, and potential clinical liability. Ultimately, the strategic imperative for health systems is to adopt a measured, phased approach, utilizing the proposed readiness framework to systematically build the necessary competencies before exposing patients to the probabilistic outputs of generative artificial intelligence.

5. Conclusion and Future Directions

5.1 Summary of Contributions

This research establishes a vital theoretical and operational framework for addressing one of the most pressing challenges in contemporary health informatics: the safe, auditable institutional integration of generative large language models. By explicitly shifting the focus away from the performance metrics of the algorithms themselves and toward the infrastructural, governance, and workflow readiness of the adopting organizations, this paper provides a necessary corrective to the pervasive techno-optimism dominating the current discourse. The multi-pillar Institution-Level Readiness Model developed herein offers healthcare administrators a concrete taxonomy for evaluating their systemic vulnerabilities. Furthermore, the emphasis on continuous auditing pipelines and dynamic evaluation metrics ensures that institutions are equipped not merely to deploy these models, but to maintain rigorous oversight of their behavior in complex, ever-evolving clinical environments. The simulated application across diverse hospital archetypes concretely demonstrates the applicability of the model while starkly highlighting the structural inequalities that must be addressed to prevent a widened digital divide in healthcare quality.

5.2 Long-term Perspectives on Clinical AI

Looking forward, the trajectory of generative artificial intelligence in medicine necessitates an ongoing evolution of these institutional readiness frameworks. As large language models transition from text-based processing to multimodal reasoning, incorporating real-time physiological data, medical imaging, and genomic sequences, the complexity of auditing will increase exponentially [20]. Health systems will need to develop entirely new paradigms of systemic oversight that can continuously evaluate multidimensional diagnostic reasoning. Furthermore, the eventual shift from human-in-the-loop copilot systems to fully autonomous clinical agents will require absolute certainty in institutional readiness and faultless auditing mechanisms. Ensuring the safety, equity, and efficacy of patient care in this rapidly approaching future depends entirely on the rigorous foundational work undertaken by healthcare institutions today to establish transparent, accountable, and highly resilient operational infrastructures.

References

- Chen, Z., Gao, P., Lee, Y., Barthelemy, J., Zhou, L., & Wang, L. (2025, June). Optimizing Efficiency and Visual-Textual Alignment for LLM-Based Radiology Report Generation. In 2025 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE.
- Xia, X., Wang, X., Li, Q., Li, N., & Li, J. (2013). Essential amino acid enriched high-protein enteral nutrition modulates insulin-like growth factor-1 system function in a rat model of trauma-hemorrhagic shock. *PLoS One*, 8(10), e77823.
- Sha, L., Wang, Y., Meng, P., Deng, Y., Chen, T., Zhang, X., ... & Xu, Q. (2025). Pharmacological inhibition of PSPH reduces serine levels and epileptic seizures. *Nature Chemical Biology*, 21(11), 1742-1753.

- Zhao, K., Li, G., Li, C., Jiang, T., Wang, W., & Yang, Z. (2026). Identification Markers for *Salvia miltiorrhiza* and Its Close Relatives Based on Cell Wall Component Characteristics. *Engineered Science*, 40, 2122.
- Gao, X. L., & Wang, M. (2026). Enhancing cancer immunotherapy antibody discovery: Genetic algorithms and risk-aware protein modeling. Available at SSRN 6033537.
- Lyu, M., Torii, R., Liang, C., Zhang, X., Wang, X., Li, Q., ... & Chen, D. (2025). Predictive computational framework to provide a digital twin for personalized cardiovascular medicine. *Communications Medicine*, 5(1), 370.
- Lyu, M., Torii, R., Liang, C., Li, Q., Wang, X., Ventikos, Y., & Chen, D. (2025). In silico comparison of two non-invasive pre-procedural virtual coronary revascularisation techniques for personalised cardiovascular medicine. *Computer Methods and Programs in Biomedicine*, 109046.
- Yu, X., Zhang, Y., & Wang, H. (2025, July). Subjective Sleep Gains in Empty Nose Syndrome: A Cautionary Note on the Need for Objective Validation. In *International Forum of Allergy & Rhinology* (Vol. 15, No. 7, pp. 761-762).
- Verhoven, Bret, et al. "Attenuating Ischemia and Reperfusion Injury Using NAD⁺-Loaded Nanoparticles in Mouse Kidneys." *Transplantation Direct* 12.1 (2025): e1890.
- Yin, Li, et al. "Targeted NAD⁺ Delivery for intimal Hyperplasia and Re-endothelialization: a novel anti-restenotic therapy approach." *bioRxiv* (2024): 2024-02.
- Sinha, D., Tong, Y., Zepeda, M. A. F., Hanstad, G., Nelson, E. C., Theisen, C. O., ... & Gamm, D. M. (2023). Silica nanocapsules as a nonviral delivery platform for iPSC-RPE. *Investigative Ophthalmology & Visual Science*, 64(8), 774-774.
- Kabra, Meha, et al. "Nonviral base editing of KCNJ13 mutation preserves vision in a model of inherited retinal channelopathy." *The Journal of Clinical Investigation* 133.19 (2023).
- Xia, F., Li, B., An, B., Zachman, M. J., Xie, X., Liu, Y., ... & Cheng, Y. (2024). Cooperative Atomically Dispersed Fe–N₄ and Sn–N_x Moieties for Durable and More Active Oxygen Electroreduction in Fuel Cells. *Journal of the American Chemical Society*, 146(49), 33569-33578.
- Hoy Jr, D., Kuang, Y., Asres, A., Liu, C. H., Dormidontova, E., & Nieh, M. P. (2025). Spontaneous Graphene/Lipid Bicellar Co-assembly (NANO2-Graphene): Experiments and Computer Simulations. *ACS Applied Materials & Interfaces*, 17(49), 66221-66234.
- Chai, L., Zhang, J., Li, H., Zheng, B., Jiang, J., Cui, C., & Jiang, L. (2020). Investigation for a multi-silique trait in *Brassica napus* by alternative splicing analysis. *PeerJ*, 8, e10135.
- Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods*, 12(10), 931–934.
- De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., et al. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24(9), 1342–1350.
- Cox, J., & Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, 26(12), 1367–1372.
- Yu, L., Li, J., Hong, J., Takashima, Y., Fujimoto, N., Nakajima, M., ... & Liu, L. (2018). Low cell-matrix adhesion reveals two subtypes of human pluripotent stem cells. *Stem cell reports*, 11(1), 142-156.
- Wang, Z., Kim, E. H., Oh, S. K., Pedrycz, W., Fu, Z., & Yoon, J. H. (2024). Reinforced fuzzy-rule-based neural networks realized through streamlined feature selection strategy and fuzzy clustering with distance variation. *IEEE Transactions on Fuzzy Systems*, 32(10), 5674-5686.