engineering proceedings

MDPI

Proceeding Paper

Designing an AI Agent System to Execute Biodesign Debate Process †

Ya Chuan Chen [1,2,*], Shih-Huan Lin [2], Ke-Wei Chen [2] and Hsiang-Wei Hu [1,2,*]

[1] Information and Communications Research Laboratories, Industrial Technology Research Institute (ITRI), Hsinchu 31040, Taiwan
[2] Department of Artificial Intelligence in Healthcare, International Academia of Biomedical Innovation Technology, Reno, NV 89507, USA; sean040@gmail.com (S.-H.L.); kilong31442@gmail.com (K.-W.C.)
* Correspondence: joycechen108@gmail.com (Y.C.C.); hw.hsiang.wei@gmail.com (H.-W.H.)
† Presented at the 7th Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability 2025 (ECBIOS 2025), Kaohsiung, Taiwan, 23–25 October 2025.

Abstract

Early-stage healthcare innovation depends on systematic unmet need discovery, a process constrained by time and multidisciplinary coordination. We developed BioDesign Agent, a multi-agent debate framework built on LangGraph to augment the identify phase of design thinking. The system assigns expert roles, clinical, engineering, human factors, regulatory, business, intellectual property, and patient access, to digital agents engaging in structured debate and scoring. Applied to antimicrobial resistance risk prediction, the agent surfaced diverse perspectives, refined need statements, and produced prioritized evaluations. Multi-agent debate yielded more differentiation, richer trade-off analysis, and more actionable insights compared with ChatGPT oss:20b only baselines, demonstrating how structured AI-assisted debate can accelerate healthcare need discovery and complement human-driven biodesign with scalable front-end innovation support.

Keywords: design thinking; multi-agent systems; structured debate; healthcare innovation; need discovery; biodesign

1. Introduction

Healthcare innovation often struggles with the persistent divide between clinical insight and engineering translation. Clinicians are adept at identifying pain points but lack pathways to convert them into technically feasible solutions, while engineers excel at design but may overlook the lived realities of patients and providers. This disconnect contributes to high failure rates in translational research, where advanced technologies fail to achieve meaningful adoption [1]. The Stanford Biodesign framework addresses this challenge by structuring innovation into three phases, Identify, Invent, and Implement, ensuring that innovators begin with the right problem before pursuing solutions [1]. However, the Identify phase remains resource-intensive, requiring multidisciplinary workshops that are difficult to scale across institutions.

Recent advances in large language models (LLMs) [2] and multi-agent systems (MAS) [3] offer new engineering pathways to augment this process. MAS frameworks simulate team-based reasoning, enabling diverse perspectives to converge on actionable insights [4,5]. Structured debate protocols, in particular, mirror real-world expert deliberation by introducing critique, revision, and consensus-building [6]. This paper introduces the BioDesign Multi-Agent Debate System, an engineering-driven framework built on

Check for updates

Academic Editors: Teen-Hang Meen, Cheng-Yi Chen and Cheng-Fu Yang

Published: 16 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

LangGraph [7], designed to augment unmet need discovery in healthcare. By orchestrating specialized agents across clinical, technical, regulatory, and socio-economic domains, the system provides transparent, auditable, and reproducible outputs that complement human-driven Biodesign.

The technical foundation of this system is a directed acyclic graph structure within LangGraph [7] to manage the state of a Need Statement as it evolves through iterative refinement. Different from linear prompting, this architecture allows for cyclic feedback loops where a Clinical Agent first characterizes a pathology or workflow inefficiency. This qualitative input is processed by a Technical Agent, which assesses the state-of-the-art and identifies engineering bottlenecks, such as power density requirements or biocompatibility constraints. This interplay ensures that the "Invent" phase is grounded in physical and computational reality from the earliest stages of problem definition.

To mitigate the risks of hallucination or overly optimistic technical forecasting, the system employs a multi-turn debate protocol [8,9]. In this phase, a Regulatory Agent and a Socio-Economic Agent act as adversarial critics, challenging the clinical-technical consensus by introducing constraints related to the Food and Drug Administration pathways, reimbursement codes, such as current procedural terminology and Healthcare Common Procedure Coding System, and health equity impact. This engineering of synthetic friction enables the agents to revise the initial need statement until it satisfies a multi-dimensional objective function: clinical impact, technical viability, and market sustainability.

From a systems engineering perspective, MAS is used to transform the Identify phase from a boutique, human-intensive activity into a scalable computational workflow. By capturing the reasoning traces of each agent, the framework generates a Discovery Log that serves as an auditable record of the innovation process. This reproducibility is critical for institutional knowledge transfer and reduces the cognitive load on human experts to change their focus from raw data synthesis to high-level strategic validation. The BioDesign Multi-Agent Debate System serves as a force multiplier, aligning technical specifications with clinical imperatives to increase the probability of successful translational outcomes.

2. Related Work

Biodesign emphasizes unmet need discovery, iterative invention, and structured implementation, ensuring that innovators focus on clinically relevant problems before pursuing technical solutions [1]. In such biodesigns, AI has been applied to augment innovation. For example, Google's AI Co-Scientist demonstrates the potential of LLMs to generate hypotheses in biomedical research, highlighting how computational reasoning can accelerate early-stage discovery [2]. Similarly, AlphaFold has transformed protein science by enabling highly accurate structure prediction, underscoring AI's capacity to address complex scientific challenges previously constrained by experimental limitations [8].

In healthcare, the MAS is regarded as a promising tool to emulate real-world teamwork. A Self-Evolving LLM-based Multi-Agent Framework for Multi-Disciplinary Team Medical Consultation is used to simulate multidisciplinary consultations by orchestrating specialized agents across clinical domains [4], while frameworks such as Medical Decision-making Agents and a Medical Multi-Agent Framework for Automating Appointment Scheduling Based on Patient–Provider Profile Resource Matching coordinate reasoning in acute care contexts, demonstrating how MAS can replicate the collaborative dynamics of hospital teams [4]. These systems illustrate the engineering potential of MAS to reduce bottlenecks in decision-making and improve coordination across diverse stakeholders.

Beyond healthcare, diverse multi-agent debate protocols have been introduced to improve reasoning by incorporating role diversity [5]. This approach enables agents to critique, refine, and converge on solutions, mirroring the structured deliberation of expert

panels. Self-consistency methods enhance reasoning accuracy by aggregating multiple independent reasoning paths into a unified output [6]. Reflexion and constitutional AI extend this paradigm by embedding self-reflection and AI feedback mechanisms, aligning outputs with human values and improving robustness [9].

Recent developments in multi-agent orchestration frameworks, such as LangGraph [7], provide engineering infrastructures for scalable collaboration among agents. These systems present traceability, auditability, and modularity, ensuring that debate outputs can be transparently evaluated and reproduced. Collectively, these advances demonstrate that multi-agent debate is not only a theoretical construct but also a practical engineering methodology for augmenting biodesign processes. By integrating structured debate, role specialization, and self-consistency, MAS frameworks can bridge the clinical–engineering divide and accelerate unmet need discovery in healthcare innovation.

3. System Architecture

3.1. System Overview

The BioDesign Multi-Agent Debate System is built on LangGraph and operates as a network of collaborating agents. The system receives a broad problem statement or unmet clinical need from the healthcare domain. Multiple expert roles are instantiated, each with dedicated context specifications establishing domain perspectives. Based on these configurations, the system orchestrates structured debate rounds across iterations, ensuring both reasoning diversity and discussion consistency. Its design principles include (1) role specialization aligned to clinical decisions; (2) traceability and auditability retaining assumptions, evidence, trade-offs, and versioned artifacts; (3) scenario portability across hospitals and regions; and (4) offline operability using local LLMs to satisfy healthcare data-protection requirements (Figure 1).

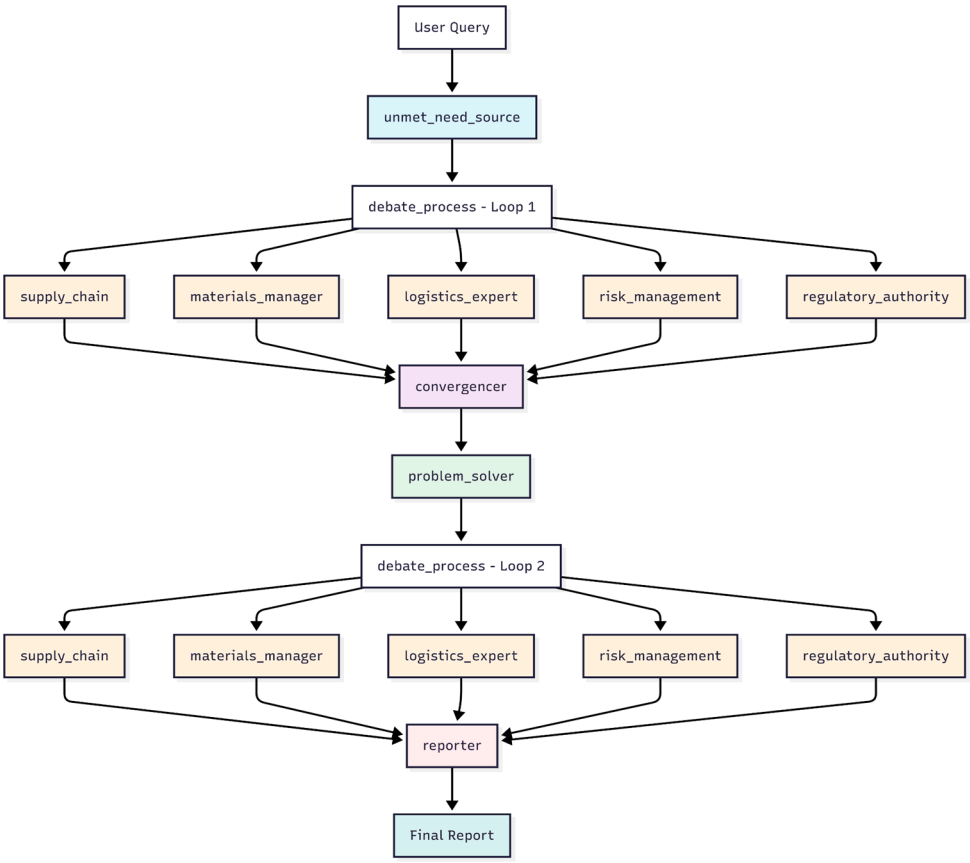


Figure 1. Architecture of the biodesign multi-agent debate system.

3.2. Agent Roles and Orchestration Workflow

The system implements specialized agents mirroring real multidisciplinary teams: clinical expert (assesses medical necessity and patient impact), engineering expert (evaluates technical feasibility), regulatory expert (reviews compliance pathways), business expert (analyzes market viability and return on investment), intellectual property expert (identifies patent landscapes), human factors expert (evaluates usability and workflow integration), and patient advocate (represents patient perspectives and access barriers). The orchestration workflow presents the following: (1) sequential initiation: the needs-source agent frames the scenario and priority tiers; (2) loop–parallel debate (round 1): domain experts post positions and analyses to a shared blackboard; (3) sequential convergence: the convergencer builds a multi-criteria matrix and consensus summary; (4) sequential option formation: the problem-solver composes 3–5 implementable options with resources, risks, and phased roadmaps; (5) loop–parallel review (round 2): a targeted re-debate of converged options; and (6) sequential reporting: the reporter produces auditable outputs, including an executive brief and a full report.

3.3. Debate Protocol and Convergence

The debate protocol consists of the following: position (domain-specific stances on clinical impact, technical feasibility, logistics, regulatory viability), critique (cross-agent challenges on risks, dependencies, boundary conditions with 3–5 configurable rounds), revise (update assumptions and options based on critiques), and optional Delphi (re-rate disputed items to reduce variance). Convergence uses an optional healthcare multi-criteria decision-making matrix weighting system impact (40%), implementability (25%), risk control (20%), innovation (10%), and cost–benefit (5%), with $\pm$15% weight sensitivity and consistency check (consistency ratio < 0.1). Final deliverables include an executive summary, full technical report with data provenance and sensitivity analysis, a stakeholder communication kit with clinical usage notes and SOPs, and traceable artifacts with prompts, intermediate outputs, and evidence links per round.

4. Case Study: Antimicrobial Resistance (AMR) Risk Prediction

4.1. Background

The multi-agent debate framework was applied to a complex clinical challenge: orthopedic post-surgical infections compounded by increasing methicillin-resistant *Staphylococcus aureus* (MRSA) prevalence and delayed antimicrobial susceptibility reporting (5–7 days). This framework necessitates a multidisciplinary engineering approach encompassing diagnostics, supply chain logistics, regulatory compliance, and financial sustainability. The system's orchestrated deliberation presents the following critical, engineering-driven requirements.

- Diagnostic acceleration: Transitioning from traditional workflows to automated systems such as Polymerase Chain Reaction and Matrix-Assisted Laser Desorption/Ionization–Time of Flight Mass Spectrometry to reduce turnaround time to under six hours [8].
- Workflow integration: Deploying infection dashboards and automated sample delivery systems to streamline laboratory-to-bedside operations.
- Regulation: Strengthening Emergency Use Authorization frameworks to facilitate the rapid adoption of new diagnostics [4].

While the framework identified significant opportunities, it also highlighted systemic risks such as false positives, alert fatigue, and the lack of longitudinal evidence. To address these, patient advocates emphasized balancing technological automation with clinical

vigilance. The resulting roadmap proposes a phased implementation: starting with short-term pilots in orthopedic wards, followed by mid-term expansion to high-risk departments, and culminating in a long-term, hospital-wide infection control platform.

4.2. Multi-Agent Debate Outcomes

The debate system on biodesign synthesized perspectives from infection specialists, laboratory experts, hospital managers, regulators, and patient advocates.

The discussion highlighted several urgent requirements. Diagnostic turnaround time must be reduced to less than six hours. Standardized MRSA prevention bundles must be established in orthopedic surgery. Automated laboratory systems, such as polymerase chain reaction and matrix-assisted laser desorption/ionization–time of flight mass spectrometry, need to be implemented. Financial models and return-on-investment analyses must be designed to ensure sustainable adoption. In addition, regulatory pathways for Emergency Use Authorization of rapid diagnostics should be strengthened.

The debate also identified key risks. Financial sustainability remains a challenge, with upfront investment estimated at NT$27 million and reimbursement pathways uncertain. Technical limitations include the possibility of false positives and alert fatigue when integrating with clinical decision support systems. Clinical evidence gaps persist, as long-term studies on proposed rapid tests are insufficient. Patient safety concerns were raised regarding the risk of over-reliance on technology, which could reduce clinical vigilance.

The roadmap for implementation was structured across three phases. In the short term, rapid MRSA and ventilator-associated event testing should be piloted in orthopedic wards, accompanied by the introduction of infection dashboards. In the mid-term, sample delivery should be automated and testing expanded to other high-risk departments. In the long term, the system should be scaled to a hospital-wide infection control platform incorporating AI-driven prediction.

4.3. Critical Assessment and Comparative Analysis

The debate revealed multi-stakeholder integration, phased implementation, and technology adoption potential. It also highlighted limitations such as financial and workflow disruption risks, a lack of comparative-effectiveness evidence, and potential regulatory bottlenecks. These findings align with known pitfalls in healthcare innovation where engineers often underestimate clinical workflow realities, and clinicians lack engineering translation pathways.

In the ChatGPT-only baseline, the system generated a concise but optimistic plan that did not adequately highlight evidence gaps or financial risks. By contrast, the multi-agent debate surfaced cross-domain risks, including reimbursement delays, alert fatigue, and regulatory compliance challenges. This comparison demonstrates the added discriminative value of structured debate, which enables a more transparent assessment of risks and benefits. Before large-scale adoption, a controlled pilot study must be conducted with rigorous metrics, including infection rates, diagnostic turnaround times, cost savings, and potential workflow disruption. Traditional infection control practices, such as hand hygiene and surgical site preparation, should be integrated with emerging rapid diagnostic technologies to ensure comprehensive protection. Regulatory engagement must be established early to align Emergency Use Authorization and reimbursement pathways. Finally, sensitivity analyses should be incorporated into financial models to account for the impact of false positives and training costs.

https://doi.org/10.3390/engproc2026129029

5. Results and Discussion

The multi-agent deliberation emphasized the role of predictive, data-driven architectures that integrate patient longitudinal history with real-time epidemiological streams to enable early-stage, precision antibiotic guidance. From the perspectives of bioengineering, a case study on post-surgical infections, including MRSA and ventilator-associated events, demonstrated that rapid diagnostic integration is essential for improving clinical outcomes. However, the engineering transition from traditional workflows to high-throughput diagnostic systems faces significant barriers, including high initial capital expenditure, the risk of disrupting established clinical laboratory workflows, and a current lack of long-term evidence regarding sustained technical efficacy.

To address such challenges, a robust implementation roadmap is proposed in this study, focusing on phased technical validation and financial sustainability. By executing staged pilot studies, institutions can verify the technical feasibility of new diagnostic hardware before full-scale deployment, while the application of conservative ROI models helps address the complexities of medical reimbursement and upfront costs (NT$27 million). Furthermore, the debate underscored the necessity of a hybrid operational model where foundational infection control engineering, such as optimized surgical site preparation and sterilized fluidics, operates in parallel with advanced sensing technologies to ensure redundant layers of patient safety.

To evaluate the computational value of the BioDesign Multi-Agent Debate System, a comparative performance benchmark was conducted against a single-agent baseline utilizing GPT-5 across five high-impact engineering domains, including antimicrobial resistance (AMR) clinical decision support. The results indicated that while single-agent systems often produced optimistic but generic recommendations, such as escalating to stronger antibiotics only upon patient deterioration without risk stratification, the multi-agent framework enabled structured, cross-domain reasoning to design early-warning risk models using Electronic Health Records (EHRs) and epidemiology data. This multi-agent collaboration enabled personalized, preventive solutions and tiered stewardship strategies that move beyond trial-and-error escalation, successfully identifying hidden system dependencies (Table 1).

Table 1. Comparative evaluation: multi-agent debate system and single-agent baseline.

Case Study	Single-Agent (GPT-5) Output	Multi-Agent Debate Output	Comparison
AMR therapy	Recommended escalation to stronger antibiotics upon deterioration, without risk stratification.	Designed early-warning risk models using EHR + epidemiology data, integrated rapid diagnostics, and tiered stewardship strategies.	Multi-agent collaboration enabled personalized, preventive solutions beyond trial-and-error escalation.

From a bioengineering systems perspective, such results highlight the need for interoperable data pipelines, standardized diagnostic protocols, and adaptive monitoring platforms to support the next generation of healthcare technology [6,7]. By embedding these structured debate protocols into hospital IT architectures, the BioDesign framework transforms resource-intensive human workshops into scalable, reproducible computational workflows. The framework replicates need-finding exercises across diverse clinical departments and geographical regions, contributing to the development of national-level innovation infrastructures. Through this engineering-driven approach, the framework systematically uncovers critical dependencies, such as supply chain fragility or specific regulatory hurdles, that are frequently neglected in traditional, human-centric innovation models [10].

6. Discussion

The BioDesign Multi-Agent Debate System demonstrates three recurring advantages over single-agent baselines by bridging the gap between clinical insight and engineering execution. By synthesizing expertise across clinical, technical, regulatory, and socio-economic dimensions, MAS provides solutions that are technically feasible and institutionally actionable. While single-agent baselines, such as GPT-5, tend to produce optimistic but generic recommendations, the MAS framework generates structured outputs, including monitoring platforms, governance matrices, and adaptive pipelines, that are directly linkable to specific evaluation metrics or regulatory requirements.

The framework's ability is critical to explore systemic vulnerabilities and hidden dependencies that are often invisible to single-agent systems or traditional human-centric workshops. These include complex bottlenecks such as raw material concentration in supply chains, reimbursement delays, and patient-specific antimicrobial resistance risk factors. By exposing these variables, the system expands the horizon of preventive interventions and aligns technical proposals with rigorous policy frameworks.

The scalability of this engineering-driven protocol shows significant potential across various sectors. In healthcare, it can be embedded into hospital IT systems to replicate need-finding exercises across departments, creating a pathway toward national-level innovation infrastructures. Beyond clinical settings, the framework's utility extends to public policy and urban planning for environmental risk management, adaptive learning platforms in education to address learning poverty, and financial regulation by translating abstract mandates into concrete audit trails. Furthermore, it serves as a force multiplier in scientific discovery by accelerating hypothesis generation and research prioritization to complement traditional expert panels.

Despite the demonstrated feasibility of the system, several engineering and operational limitations exist. Current validation is mainly focused on a single case study. Therefore, broader testing across diverse clinical contexts and larger sample sizes is required to ensure generalizability. From a technical perspective, the computational costs associated with multiple debates restrict real-time applicability in resource-constrained settings. Additionally, the definitions of agent roles require ongoing refinement based on continuous feedback from domain experts to maintain high-fidelity reasoning.

Therefore, it is necessary to develop adaptive debate protocols to dynamically adjust the number of deliberations based on real-time disagreement levels and reduce computational overhead. There is also a need to explore hybrid human–AI panels that integrate clinicians and engineers into specific roles within the graph, ensuring a contextual foundation. For standardized adoption, evaluation benchmarks and metrics are necessary to quantify debate quality and practical utility. Finally, extensive deployment studies within national health systems and municipal governments are necessary to assess the long-term cost-effectiveness, user acceptance, and the necessity for interoperable data pipelines and standardized diagnostic protocols.

7. Conclusions

We developed a BioDesign Multi-Agent Debate System, an engineering-driven system built on LangGraph designed to augment the Identify phase of healthcare innovation. By integrating specialized, domain-specific agents through structured debate rounds, the system generates, critiques, and prioritizes unmet clinical needs with significantly greater depth, rigor, and reasoning diversity than single-agent baselines.

When applied to AMR risk prediction, the system presents complex cross-domain risks and highlights critical evidence gaps. The resulting outputs provided actionable implementation roadmaps that effectively balance technical feasibility with institutional

and regulatory constraints. By embedding core engineering principles, specifically traceability, auditability, and interoperability, the system serves as a robust foundation for evidence-based decision-making that directly complements human expertise and accelerates front-end innovation.

The system also demonstrates that computationally assisted deliberation enhances the robustness of the discovery process and creates a scalable infrastructure for innovation. Beyond clinical applications, this methodology holds significant promise as a paradigm for addressing high-stakes challenges in broader domains, including public policy, education, and financial regulation. The developed BioDesign Multi-Agent Debate System functions as a transparent and reproducible engine for identifying and refining the right problems before pursuing technical solutions.

Author Contributions: Conceptualization, H.-W.H.; methodology and study design, Y.C.C.; software development and testing, S.-H.L.; clinical validation, K.-W.C.; writing, original draft preparation, Y.C.C.; writing, review and editing, Y.C.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zenios, S.; Makower, J.; Yock, P. *Biodesign: The Process of Innovating Medical Technologies*; Cambridge University Press: Cambridge, UK, 2015.
2. Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S.S.; Wei, J.; Chung, H.W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. Large language models encode clinical knowledge. *Nature* **2023**, *620*, 172–180. [CrossRef] [PubMed]
3. Wooldridge, M. *An Introduction to MultiAgent Systems*, 2nd ed.; John Wiley & Sons: Chichester, UK, 2009.
4. Chen, K.; Li, X.; Yang, T.; Wang, H.; Dong, W.; Gao, Y. MDTeamGPT: A Self-Evolving LLM-based Multi-Agent Framework for Multi-Disciplinary Team Medical Consultation. *arXiv* **2025**, arXiv:2503.13856.
5. Liu, Y.; Cao, J.; Li, Z.; He, R.; Tan, T. Breaking Mental Set to Improve Reasoning through Diverse Multi-Agent Debate. In Proceedings of the Thirteenth International Conference on Learning Representations (ICLR 2025), Singapore, 24–28 April 2025.
6. Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; Zhou, D. Self-Consistency Improves Chain-of-Thought Reasoning in Language Models. *arXiv* **2022**, arXiv:2203.11171.
7. LangGraph: Multi-Agent LLM Workflows. Available online: https://blog.langchain.com/langgraph-multi-agent-workflows/ (accessed on 29 December 2025).
8. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [CrossRef] [PubMed]
9. Shinn, N.; Cassano, F.; Berman, E.; Gopinath, A.; Narasimhan, K.; Yao, S. Reflexion: Language Agents with Verbal Reinforcement Learning. *arXiv* **2023**, arXiv:2303.11366. [CrossRef]
10. Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Li, B.; Zhu, E.; Jiang, L.; Zhang, X.; Zhang, S.; Liu, J.; et al. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. *arXiv* **2023**, arXiv:2308.08155.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.