Assurance in an AI Dual-use Era: A Cyberbiosecurity Workflow Vision

Celisha Oscar[1], Lucas Potter[2] and Xavier-Lewis Palmer[2]
[1]Howard University, Washington DC, USA
[2]BIOSView Labs, Ohio, USA

Celisha.oscar@bison.howard.edu

Abstract: Artificial intelligence (AI) has become a central component of modern drug discovery, accelerating the exploration of chemical and biological design spaces using evolving computational workflows. As these systems mature, AI-assisted drug design is recognized as a tool with dual-use capabilities, raising concerns related to misuse, repurposing, and unintended system behavior. Current governance approaches largely focus on deployment settings, model performance, and intended applications, hindering visibility into the internal parameters that can conflict with system-level assurance, governance, and oversight. In this work, we propose a parameter-level assurance perspective for AI drug design systems. Rather than treating models as static artifacts, we frame key model parameters as security-relevant interfaces that warrant improved oversight. A prototype cyberbiosecurity assurance workflow centered on risk-tiered classification, adversarial stress testing, and continuous monitoring is introduced. This is proposed as a way to begin addressing the challenges involving responsible deployment and threat models.

Keywords: Drug discovery, Dual-use, Biocybersecurity, Cyberbiosecurity, System-level governance

1. Introduction

Artificial intelligence (AI) and machine learning (ML) have revolutionized drug discovery by streamlining the identification of novel targets and reducing pharmaceutical development costs and attrition rates (Dharmasivam et al., 2026; Fu et al., 2025; Riemer et al., 2026). However, the efficiencies gained through continuous learning and the exploration of expansive chemical landscapes introduce significant governance challenges, as the inherent fluidity of these models complicates oversight. Emerging scholarship underscores the risk of scientific tools functioning as adversarial vectors, particularly in chemical, biological, radiological, and nuclear (CBRN) contexts (OpenAI, 2024). Empirical evidence confirms this fragility: LLMs have provided granular procedural guidance for laboratory tasks, lowering barriers to dual-use knowledge (Soice et al., 2023; Das et al, 2026; Palmer et al, 2026). More critically, generative models have inverted reward signals to produce nerve agent candidates or synthesized structures analogous to ricin, diphtheria toxin, and sarin, lacking adequate safety guardrails (Urbina et al., 2022; Hattoh et al., 2025). The lack of robust, continuous governance exacerbates these biomanufacturing vulnerabilities downstream. Moving beyond reactive, post-hoc oversight, we propose a novel architecture employing proactive risk sensors for real-time dual-use mitigation, shifting the paradigm from static review to dynamic, parameter-level intervention.

2. Method and Vision

Our methodology is predicated on dual-use mitigation strategies that extend AI governance beyond post-hoc detection by treating model behavior as a dynamic interface rather than a static artifact. To implement this, we propose a layered assurance framework utilizing two parallel risk sensors: Qwen3Guard, a transformer-based safety classifier (Venegas 2026), and PyroSage, a graph neural network pre-trained on acute toxicity datasets (PyroSage 2025). Furthermore, Tanimoto similarity metrics are employed to identify out-of-distribution (OOD) structures that share features with known toxins within the T3DB dataset (Lim et al. 2010; Wishart et al. 2015). Figure 1 illustrates this system-level governance framework, wherein Qwen3Guard evaluates natural language inputs, while PyroSage analyzes SMILES strings to predict LD50 values (mg/kg, rat oral) and map them against Globally Harmonized System (GHS) acute oral toxicity classes (1–5). Having established this dual-path architecture, we elect to proceed to evaluate its efficacy through scenario-based stress testing, analyzing how these integrated sensors function both independently and in concert to mitigate potential dual-use threats.

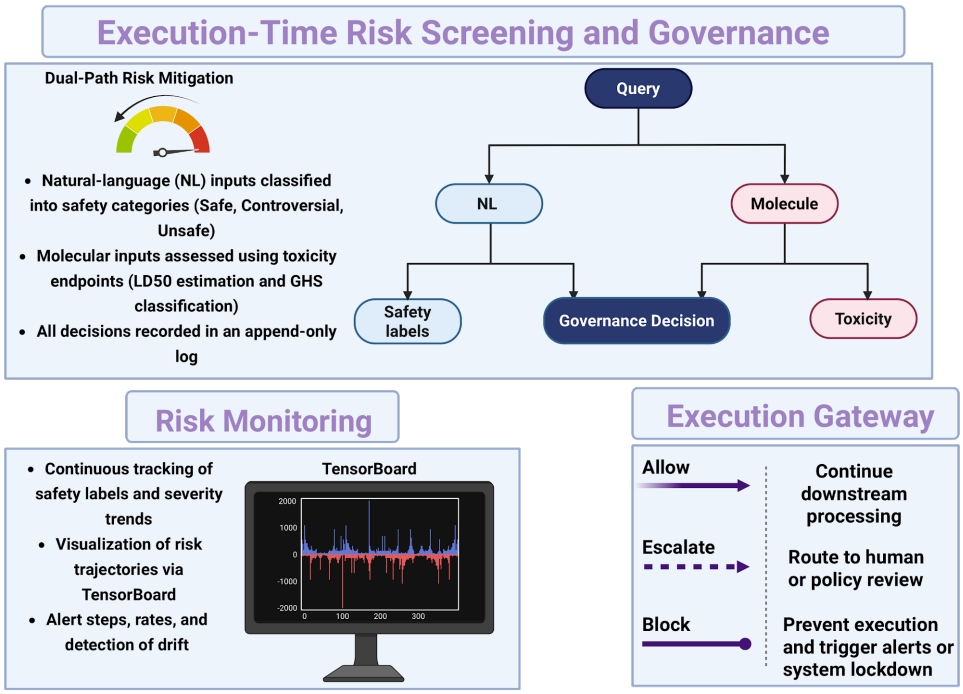

Figure 1: illustration of the framework and execution gateway

Qwen3Guard (Qwen3Guard-Gen-0.6B) was selected over alternatives due to its open-weight model that produces a three-tier output without additional post-processing and designed for research laboratory dual-use contexts. Others are trained for non-scientific research framing. PyroSage-LD50 was selected for molecular toxicity prediction over ProTox 3.0, pkCSM, ADMETLab 3.0, and ADMET-AI. ProTox 3.0, pkCSM, and ADMETLab 3.0 operate as web servers requiring external API calls which are incompatible with inference-time governance. ADMET-AI was evaluated and offered the highest average rank on the TDC ADMET benchmark leaderboard but produced unresolvable dependency conflicts between its chemprop requirement and the transformers version required by Qwen3Guard in the deployment environment. PyroSage is also open-weight, deployable locally, and uses an AttentiveFP graph neural network operating natively on molecular graphs rather than descriptor vectors. T3DB was selected over ChEMBL and PubChem bioassay data because it is specifically curated for known toxins with explicit hazard classification and SMILES annotations in a single downloadable file.

3. Pre-evaluation of Vision and Discussion

Following the establishment of this dual-path architecture, we evaluated system efficacy through scenario-based stress testing. Benchmarking against a simulated reference database of five Schedule 1 compounds reveals distinct governance bottlenecks: while PyroSage maintains moderate accuracy across drug-like space (R^2=0.562 across 7,396 compounds), it struggles with GHS Class 1 compounds due to training data scarcity. Integrating PyroSage with T3DB improves CWA blocking performance from 2/10 to 7/10, quantifying previously theoretical performance gaps (see Figure 3). Complementing this, CUSUM monitoring captures toxicity accumulation patterns. System-wide metrics including mispredicted LD50, T3DB match status, and final governance decisions are audited in real-time via TensorBoard, satisfying foundational requirements for responsible AI deployment (Raji et al., 2020). Figure 2 illustrates these findings, highlighting distinct outcomes across varying input risk signals.

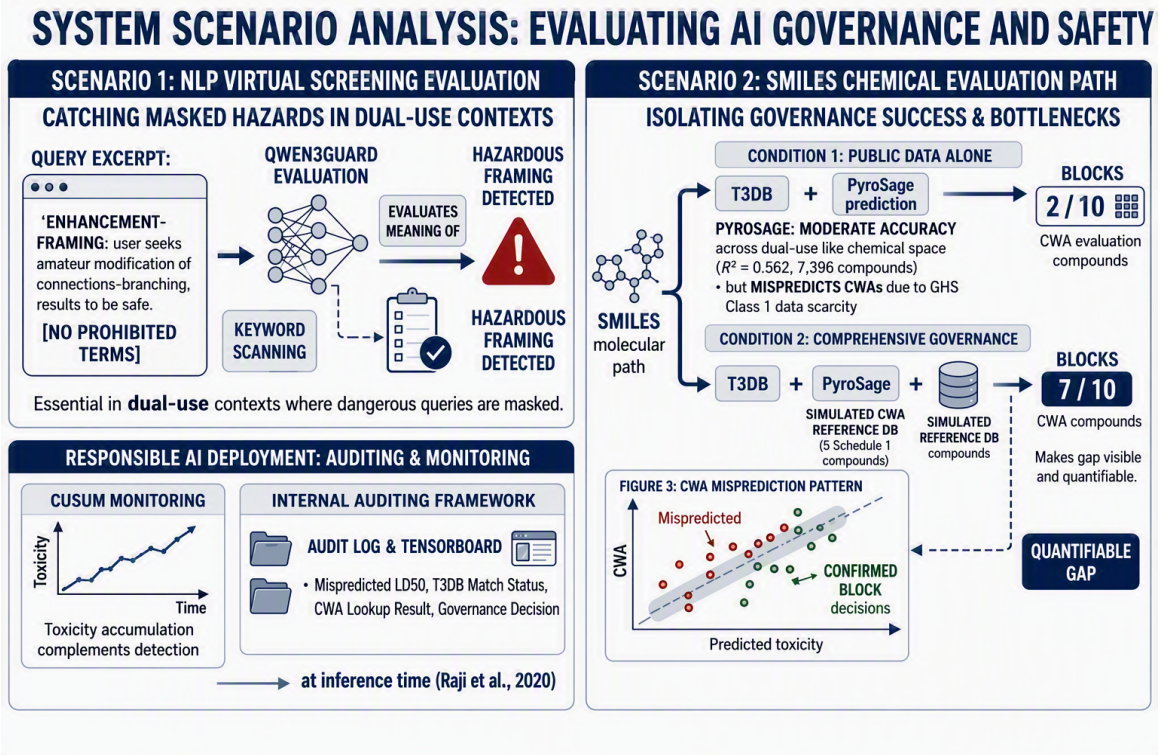

Figure 2: Scenario analyses demonstrating distinct governance decisions across different combinations of input type and risk signal

Beyond immediate risk mitigation, the broader implication of this layered assurance model is the fundamental shift toward verifiable safety in autonomous scientific research. By documenting granular governance decisions alongside toxicity trajectories, this architecture transforms static compliance checklists into dynamic, observable feedback loops. This paradigm not only creates an audit trail for regulatory transparency but also establishes a foundation for "governance as code" in drug discovery, where safety thresholds adapt in real-time as new toxicological data emerges. Such a system effectively matures the safety-critical software engineering lifecycle, ensuring that as AI agents become more capable, our ability to monitor and constrain their outputs remains commensurately robust. While these mechanisms fundamentally strengthen our defensive posture, they highlight the friction between idealized architectural security and the messy, unpredictable realities of production environments. However, translating this architectural vision into a production-ready system requires navigating significant practical constraints and technical hurdles, which are detailed in the following section.

4. Limitations and Future Work

Several factors limit our findings' generalizability. First, our evaluation employs a simulated pipeline, not a production environment; real-world workflows exhibit greater query diversity and behavioral complexity than our synthetic model captures. Second, restricted access to lethal datasets (e.g., G-series nerve agents) hampers validation. While the model performs well on drug-like compounds, LD50 accuracy declines for GHS 1-2 compounds due to training data scarcity; future improvements include fine-tuning on the DSSTox dataset. Finally, protein toxins remain outside the current molecular evaluation scope, necessitating integration with ToxinPred 2.0, and rigorous benchmarking is essential for deployment. As summarized in Figure 3, addressing these constraints is critical for maturing this framework into a production-ready governance tool, which we discuss in the following section.

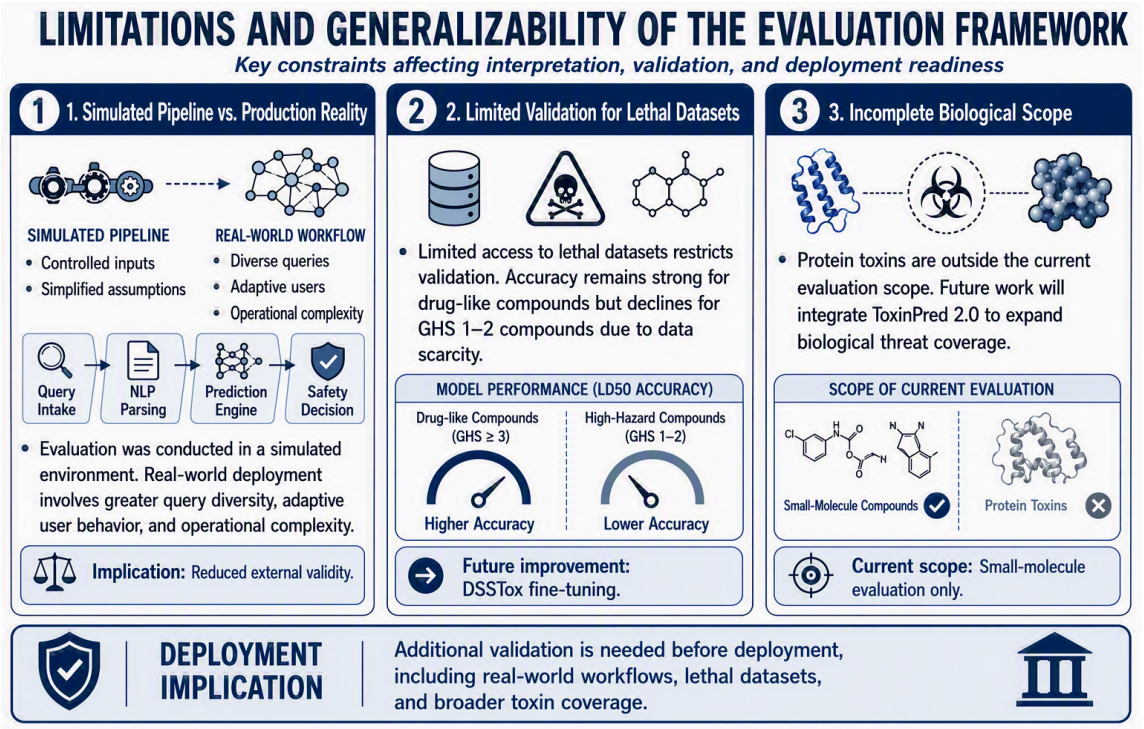

Figure 3: Limitations and Areas for Future Developement

5. Conclusion

Current regulatory frameworks govern static products, yet AI's dynamic nature creates a "living approval" paradox, complicating ongoing safety validation (Copo 2025). This disconnect between rapid integration and outdated oversight creates significant assurance blind spots. We propose a governance framework addressing the urgent need for system-level safeguards against hazardous outputs. Our model achieves four objectives: (i) enforcing inference-time policy across diverse queries, (ii) preventing harmful requests, (iii) detecting OOD inputs, and (iv) ensuring auditable monitoring of behaviors and decisions.

Ethics Declaration: All work was conducted with the purpose and intent to observe system behavior. Ethical clearance was not required for this work.

AI Declaration: The preparation of this work involved Claude 3.5 Sonnet to construct code, Gemini and ChatGPT for developing the figures, as well as tools within Google Docs to re-organize pre-written text. Model outputs were reviewed and available upon request.

References:

AeIqbtat, E.H., Roope, R., Chikersal, P., Bogen, J.M. and Esvelt, K.M. (2023) 'Can large language models democratize access to dual-use biotechnology?', *arXiv*. https://arxiv.org/abs/2306.03809

Chen, F. and Chen, Q., 2025. The future of pharmaceuticals: Artificial intelligence in drug discovery and development. Journal of Pharmaceutical Analysis, 15(8), p.101248. https://doi.org/10.1016/j.jpha.2025.101248

Copo, J., 2025. AI in drug discovery: a regulatory tightrope walk. International Bar Association, 3 December. Available at: https://www.ibanet.org/ai-drug-discovery-regulatory [Accessed 25 June 2026].

Das, P., Chakraborty, S., Narula, S., Potter, L., Palmer, X.-L., Rana, P., Takabi, D. and Ghasemigol, M. (2026) BioShield: A Context-Aware Firewall for Securing Bio-LLMs. arXiv:2603.22612 [cs.CR]

Dharmasivam, M., Kaya, B., Akinware, A. et al., 2025. Leading artificial intelligence–driven drug discovery platforms: 2025 landscape and global outlook. Pharmacological Reviews, 78(1), p.100102. https://doi.org/10.1124/pharmrev.25.07511-8

DiEuliis, D., Lutes, C.D. and Giordano, J., 2018. Biodata risks and synthetic biology: A critical juncture. Journal of Bioterrorism & Biodefense, 9(159), p.2.

Gazuzect, A.M., 2019. The national security implications of cyberbiosecurity. Frontiers in Bioengineering and Biotechnology, 7, p.51. https://doi.org/10.3389/fbioe.2019.00051

George, A.M., 2019. The national security implications of cyberbiosecurity. Frontiers in Bioengineering and Biotechnology, 7, p.51. https://doi.org/10.3389/fbioe.2019.00051

Gillum, D., Carrera, L.A.O., Mendoza, I.A., Bates, P., Bowens, D., Jetson, Z., Maldonado, J., Mancini, C., Miraldi, M., Moritz, R. and O'Donnell, M., 2018. The 2017 Arizona biosecurity workshop: An open dialogue about biosecurity. Applied Biosafety, 23(4), pp.233–241. https://doi.org/10.1177/1535676018796793

Greenbaum, D. (ed.), 2023. Cyberbiosecurity: A new field to deal with emerging threats. Cham: Springer Nature.

Hattoh, G., Ayensu, J., Ofori, N.P., Eshun, S. and Akogo, D. (2025) 'Can large language models design biological weapons? Evaluating Moremi Bio', *arXiv*. https://arxiv.org/abs/2505.17154

Kiteworks, 2025. AI data security and compliance risk report: Industry study reveals 83% of organizations operate without basic controls. Available at: https://www.kiteworks.com/sites/default/files/resources/kiteworks-report-ai-data-security-and-compliance.pdf [Accessed 25 June 2026].

Lim, E., Pon, A., Djoumbou, Y., Knox, C., Shrivastava, S., Guo, A.C., Neveu, V. and Wishart, D.S. (2010) 'T3DB: a comprehensively annotated database of common toxins and their targets', *Nucleic Acids Research*, 38(Database issue), pp. D781–D786. https://doi.org/10.1093/nar/gkp934

Mantle, J.L., Rammohan, J., Romantseva, E.F., Welch, J.T., Kauffman, L.R., McCarthy, J., Schiel, J., Baker, J.C., Strychalski, E.A., Rogers, K.C. and Lee, K.H., 2019. Cyberbiosecurity for biopharmaceutical products. Frontiers in Bioengineering and Biotechnology, 7, p.116. https://doi.org/10.3389/fbioe.2019.00116

Muratov, E.N., Bajorath, J., Sheridan, R.P., Tetko, I.V., Filimonov, D., Poroikov, V. and Tropsha, A. (2020) 'QSAR without borders', *Chemical Society Reviews*, 49(11), pp. 3525–3564. https://doi.org/10.1039/D0CS00098A

Murch, R.S., So, W.K., Buchholz, W.G., Raman, S. and Peccoud, J., 2018. Cyberbiosecurity: An emerging new discipline to help safeguard the bioeconomy. Frontiers in Bioengineering and Biotechnology, 6, p.39. https://doi.org/10.3389/fbioe.2018.00039

OpenAI, 2024. OpenAI o1 system card. Available at: https://cdn.openai.com/o1-system-card-20240917.pdf [Accessed 25 April 2026].

Palmer, X.-L., Potter, L., Lesaja, S., Karathanasis, S. and Ghasemigol, M. (2026) 'Biosecure-LLM Framework: Protecting LLMs from Cyberbiosecurity Threats and the Case for Independent AI Safety Governance', International Conference on Cyber Warfare and Security, 21(1), pp. 743–750. doi:10.34190/iccws.21.1.4524.

Pauwels, E. and Denton, S.W., 2018. The Internet of Bodies: Life and death in the age of AI. California Western Law Review, 55, p.221.

Potter, L. and Palmer, X.-L., 2023. Mission-aware differences in cyberbiosecurity and biocybersecurity policies: Prevention, detection, and elimination. In: D. Greenbaum (ed.) Cyberbiosecurity: A new field to deal with emerging threats. Cham: Springer International Publishing, pp.37–69.

Puzis, R., Farbiash, D., Brodt, O., Elovici, Y. and Greenbaum, D., 2020. Increased cyber-biosecurity for DNA synthesis. Nature Biotechnology, 38(12), pp.1379–1381. https://doi.org/10.1038/s41587-020-00735-1

UPCI NTUA, 2025. Pyrosage LD50_Zhu AttentiveFP Model. Hugging Face. Available at: https://huggingface.co/upci-ntua/pyrosage-ld50_zhu-attentivefp [Accessed 22 June 2026].

Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D. and Barnes, P. (2020) 'Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing', *Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency*, pp. 33–44. https://doi.org/10.1145/3351095.3372873

Riemer, A., Freund, V., Generative artificial intelligence in pharmaceutical drug development: A systematic review of time and cost efficiency across discovery, preclinical, and clinical phases. [Journal not specified].

Urbina, F., Lentzos, F., Invernizzi, C. and Ekins, S., 2022. Dual use of artificial-intelligence-powered drug discovery. Nature Machine Intelligence, 4, pp.189–191. https://doi.org/10.1038/s42256-022-00465-9

Venegas, G., 2026. Agile artificial intelligence governance: A practical approach to responsible corporate adoption. Zenodo. https://doi.org/10.5281/zenodo.18918455