Fourth Eon Biosecurity Institute

Identifying function-relevant signatures in protein models for biosecurity screening

Project Proposal | Isha Harris & Gary Abel | July 25, 2026 | fourtheon.bio

Research Stream Overview

Advances in AI capabilities for biology are challenging existing biosecurity safeguards, particularly DNA synthesis screening, which flags orders based on similarity to known sequences of concern (Baker & Church, 2024; Hunter, 2024). It remains unclear to what extent biological AI models can reliably generate functional genes and proteins outside natural sequence space, rather than merely interpolating within their training data. But the further that designed proteins depart from known biological distributions, the more likely sequence-based methods are to fail. A red-teaming study by Wittmann et al. (2025) showed that some AI-redesigned proteins can evade established screening methods, though it's uncertain whether these failures reflect limited generalization to divergent sequences, or practical constraints on false positives.

Protein language models are trained on large sets of protein sequences, and learn representations that can encode aspects of protein structure, function, and evolutionary relationships (Rives et al., 2021; Candido et al., 2026). These representations may reveal biologically relevant similarities that are not readily identifiable from direct sequence comparison alone (Liu et al., 2024), making them a promising complement to existing screening methods. If these features generalize beyond the region of protein space explored by nature, they may enable a function-based approach to screening that is more robust against AI-enabled design (Abel et al., 2026).

This research stream will investigate whether biological foundation models can address this screening gap. It explores embedding-based approaches and interpretability methods, such as sparse autoencoders, to identify learned representations that are biophysically meaningful (Simon & Zou, 2025; Candido et al., 2026), and therefore potentially useful for function-based screening. The aim is to determine whether these approaches offer practical advantages over BLAST, profile HMMs, and structural search without inflating the false-positive rate.

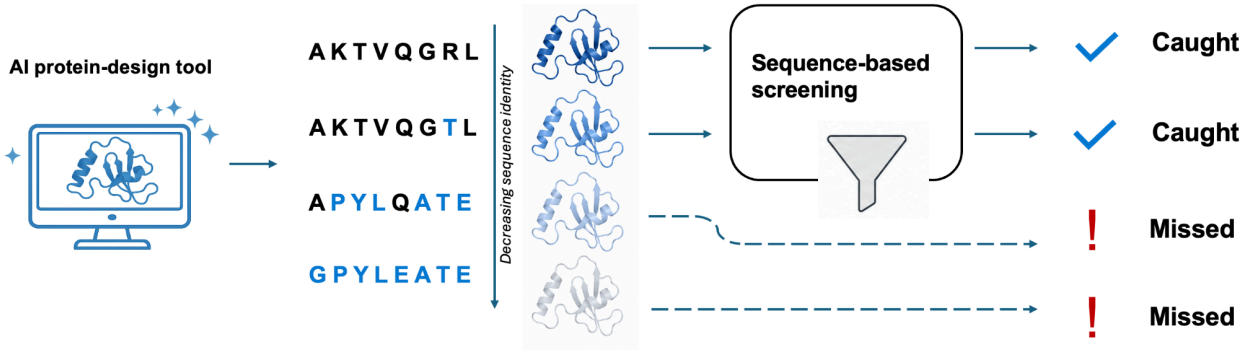

Figure 1: *Illustration of the screening gap created by structurally conserved but sequence-divergent AI-designed proteins.*

Key Objectives

1. **Identify screening-relevant representations in biological models.** Use interpretability methods to identify biophysical and biochemical properties encoded by biological models.
2. **Test whether these representations support detection of divergent proteins.** Explore whether learned representations capture shared functional properties across diverse proteins generated by AI protein design tools.
3. **Evaluate feasibility and performance against relevant baselines.** Assess whether function-based screening approaches provide performance gains beyond established sequence-based methods such as BLAST and profile HMMs. Measure impact on sensitivity and selectivity, and characterize their failure modes.

Project Selection

Fellows will work with the mentors to develop a focused research project within this broader research stream. Fellows applying to this stream are encouraged to think about potential project ideas and discuss with the mentors in advance. Projects may be empirical, methodological, or exploratory, but should clearly address a relevant hypothesis or open question related to bio model interpretability and capability measurement for biosecurity screening.

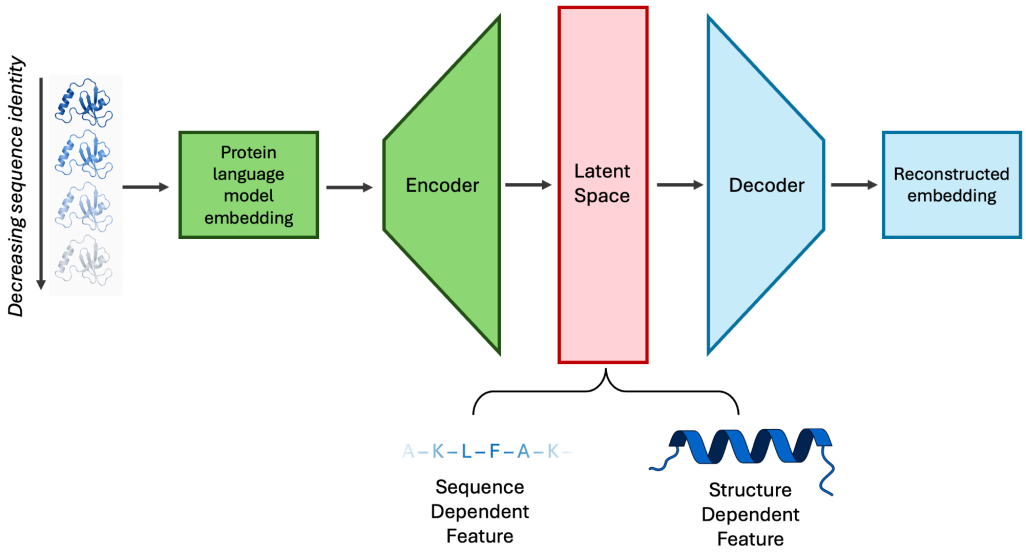

Figure 2: *Illustration of encoder–feature–decoder methods for decomposing protein language-model representations into biologically relevant features, including across highly divergent proteins*.

Responsible Research

Fourth Eon is committed to a **Safe and Responsible Research Framework** (SRRF), which includes following the recommendations of an independent Biosecurity Advisory Committee, using safe proxy systems for wet-lab experiments, responsible disclosure and managed access to sensitive data and tools, implementation of robust cybersecurity and information security controls, ongoing engagement with key external stakeholders, and an organizational culture built on safety, security and responsibility.

Note that because Fourth Eon works with sensitive methods and data, we require Fellows to sign a fellowship agreement covering confidentiality, pre-publication review of research outputs for dual-use risks, and a commitment to follow our SRRF. We are glad to provide a copy of each to prospective Fellows for review.

About the Organization

Fourth Eon Biosecurity Institute is a nonprofit organization building **AI-resilient biosecurity infrastructure**. Our first RD&D program, SCREEN, is developing function-based screening for DNA synthesis, aimed at robustly detecting hazardous sequences even when they're highly divergent from nature. Fourth Eon is built as a **permanent execution layer for biosecurity**, leveraging advances in frontier AI capabilities to research, develop and deploy adaptive safeguards for dual-use bioengineering.

References

1. Abel, G. R. Jr. et al. (2026). Beyond sequence similarity: toward function-based screening of nucleic acid synthesis. *Frontiers in Bioengineering and Biotechnology*, **14**, 1832724. https://doi.org/10.3389/fbioe.2026.1832724.
2. Baker, D. and Church, G. (2024). Protein design meets biosecurity. *Science*, **383**(6681), 349. https://doi.org/10.1126/science.ado1671.
3. Candido, S. et al. (2026). Language modeling materializes a world model of protein biology. *bioRxiv*. https://doi.org/10.64898/2026.06.03.729735.
4. Hunter, P. (2024). Security challenges by AI-assisted protein design: the ability to design proteins *in silico* could pose a new threat for biosecurity and biosafety. *EMBO Reports*, **25**, 2168–2171. https://doi.org/10.1038/s44319-024-00124-7.
5. Liu, W., Wang, Z., You, R. et al. (2024). PLMSearch: protein language model powers accurate and fast sequence search for remote homology. *Nature Communications*, **15**, 2775. https://doi.org/10.1038/s41467-024-46808-5.
6. Rives, A., Meier, J., Sercu, T. et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, **118**(15), e2016239118. https://doi.org/10.1073/pnas.2016239118.
7. Simon, E. and Zou, J. (2025). InterPLM: discovering interpretable features in protein language models via sparse autoencoders. *Nature Methods*, **22**, 2107–2117. https://doi.org/10.1038/s41592-025-02836-7.
8. Wittmann, B. J., Alexanian, T., Bartling, C. et al. (2025). Strengthening nucleic acid biosecurity screening against generative protein design tools. *Science*, **390**(6768), 82–87. https://doi.org/10.1126/science.adu8578.