# Benchmarking Biosafety in Generative Protein Design: A Stress-Test Framework for Binder Models

**Mingyang (Erik) Xu**
University of Michigan Medical School
Ann Arbor, MI, USA
xmingyan@umich.edu

## Abstract

Generative AI has transformed protein binder design, enabling rapid creation of compact proteins with high predicted foldability and affinity. Yet these advances raise biosafety concerns: current models lack refusal mechanisms, treat benign and hazardous specifications equivalently, and are easily exploitable by adversarial prompting. We introduce a governance-aware benchmark for stress-testing generative protein design models. Input specifications are stratified into three layers inspired by biosafety levels—Benign, Ambiguous, and Malicious—and evaluated along four orthogonal dimensions: Refusal, Plausibility, Safety Distance, and Adversarial Robustness. Results are reported in a Spec × Metric matrix that highlights cross-layer safety gaps without disclosing sensitive sequences. A pilot evaluation with RFdiffusion shows no refusal, plausibility scores insensitive to biosafety level, trivial robustness, and stratification only in safety distance. These findings underscore the absence of intrinsic biosafety alignment in current structural generators. Grounded in established biosafety frameworks, this benchmark provides a reproducible foundation for community standards at the intersection of generative biology, AI safety, and governance.

## 1 Introduction

Generative AI has rapidly reshaped protein design, with binder generation emerging as a flagship application that crystallizes both opportunity and risk. While the term *generative AI* is often associated with large language[1] or image models[11], we adopt the broader definition: systems that generate novel biological sequences or structures. Protein binder design models such as RFdiffusion[24], Chroma[15], and BindCraft[18]fall squarely into this category, producing *de novo* proteins conditioned on structural or sequence-level constraints.

Unlike traditional pipelines that rely on laborious screening or narrowly parameterized rational design[13], these generative systems can propose binders for arbitrary targets, accelerating therapeutic discovery, functional annotation, and synthetic biology[24, 25]. They now routinely produce compact proteins with high predicted foldability and binding to intended epitopes. Yet systematic evaluation of their *biosafety alignment* remains absent, raising critical questions about dual-use potential.

Binder generation has quickly become a canonical testbed for generative biology. A de facto pipeline integrates backbone generation (e.g., diffusion models[24]), sequence assignment (e.g., ProteinMPNN[7]), and structural validation (e.g., AlphaFold2[16]). Progress is measured largely in terms of stability, docking, or success rates[20]. But three safety-relevant questions remain unresolved: will models refuse malicious requests, will plausible outputs stay at a safe distance from known toxins, and will safeguards remain robust under adversarial reformulation? In broader AI research, stress-testing benchmarks such as HELM[17] or vBench[8] have formalized this as a *Scenario* ×

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Biosecurity Safeguards for Generative AI (NeurIPS 2025 Workshop).

*Metric* evaluation. By analogy, we define protein binder stress-testing as a function

$$f(s, m) \in \{0, 0.25, 0.5, 0.75, 1.0\}, \quad s \in \mathcal{S},\ m \in \mathcal{M},$$

where specifications $s$ (input tasks) and metrics $m$ (safety dimensions) jointly determine benchmark scores.

We address this gap with a biosafety stress-testing benchmark. The framework stratifies input specifications into three categories—Benign (BSL-1–like), Ambiguous (BSL-2–like), and Malicious (BSL-3/4 + high-consequence capabilities)—while distinguishing between therapeutic design tasks and misuse-oriented pathogen enhancement[5]. Evaluation proceeds along four orthogonal dimensions: Refusal, Plausibility, Safety Distance, and Adversarial Robustness. Results are reported as a Spec $\times$ Metric matrix that highlights different priorities across layers while avoiding disclosure of sensitive content. Finally, we present a pilot evaluation using RFdiffusion[24], chosen for its widespread adoption and demonstrated capability, to illustrate how current structural design tools reveal a fundamental dilemma: unlike sequence language models such as ESM3[12], which can in principle be trained to refuse, diffusion-based structural generators deterministically accept any input, making refusal both technically infeasible and practically ambiguous for dual-use cases.

## 2 Related Work

### 2.1 Protein Binder Design Models

Binder generation has become a focal point in *de novo* protein design, with several models establishing new standards for capability. RFdiffusion introduced diffusion-based backbone modeling for general-purpose binder design [24], Chroma provided a programmable framework for protein complexes with functional and geometric conditioning [15], and BindCraft[18] coupled AlphaFold2[16] with gradient-based optimization for automated binder discovery. These methods routinely generate compact proteins with high predicted foldability and affinity, but they focus solely on functional optimization—none incorporate mechanisms for biosafety alignment or refusal.

### 2.2 AI Safety Benchmarks

In parallel, the AI community has developed benchmarks to evaluate robustness and safety in large models. HELM formalized evaluation as a Scenario $\times$ Metric matrix, covering accuracy, calibration, efficiency, and fairness [17]. vBench and related adversarial benchmarks probe refusal behavior under jailbreak-style prompting[14]. These efforts demonstrate the value of multi-axis safety assessment, but they remain limited to language and general-purpose models.

### 2.3 BioAI Benchmarks

Benchmarking in biology has emphasized functionality rather than safety. MoleculeNet [26] and TAPE [20] standardize molecular and protein learning tasks. Recent policy discussions of "high-consequence capabilities" highlight dual-use concerns [19], yet systematic stress-testing benchmarks for biosafety are absent.

### 2.4 Gap

SafeProtein recently introduced the first red-teaming benchmark for protein language models (ESM3)[8], showing that masked sequence completion can regenerate hazardous proteins. This highlights risks in foundation models but does not address the tools most widely adopted for binder design, namely diffusion- and optimization-based structural generators such as RFdiffusion[24] and BindCraft[18]. Unlike LLMs, these models have no notion of refusal: they deterministically accept any target and hotspot as valid conditions[24]. Our benchmark directly addresses this gap, adapting the logic of AI safety evaluation to protein binder design by layering input specifications, scoring across multiple safety axes, and reporting governance-aligned results.

## 3 Benchmark Framework

### 3.1 Scope and Models

Protein binder generation has rapidly emerged as one of the most prominent applications of generative AI in the life sciences. Unlike traditional protein engineering pipelines that rely on laborious screening or rational design, generative models can directly propose candidate binders for arbitrary molecular targets, accelerating therapeutic discovery, functional annotation, and synthetic biology applications[18, 9, 23, 6]. The field has been catalyzed by recent breakthroughs in generative modeling for proteins, which now enable the *de novo* design of compact binding proteins with nanomolar to picomolar affinities[18, 4]. As a result, binder generation has become a canonical testbed for both the opportunities and the risks associated with generative AI in biotechnology.

Representative approaches include RFdiffusion [24], Chroma [15], BindCraft [18], and related SE(3)-equivariant[10] or optimization-based pipelines[3]. The benchmark proposed here is model-agnostic: it is designed to apply equally to any of these architectures. However, for proof of concept, we implement a pilot evaluation on RFdiffusion[24], given its wide adoption and demonstrated success across diverse binder design tasks.

### 3.2 Specification Stratification

To systematically evaluate the biosafety performance of binder generation models, we stratify input specifications into three layers: **Benign**, **Ambiguous**, and **Malicious**. This structure is inspired by the Biosafety Level (BSL) framework established by CDC/NIH [5], but adapted here for the evaluation of generative AI systems.

Table 1: Representative proteins at different biosafety levels (BSL) with PDB IDs.

| **BSL Level** | **Representative Proteins (PDB ID)**[2] |
|---|---|
| **BSL-1 (Benign)** | E. coli luciferase $\beta$ subunit 1BSL;<br>E. coli DNA polymerase II 1Q8I;<br>E. coli GlnK 1GNK |
| **BSL-2 (Ambiguous)** | HIV protease 1HVR;<br>HIV reverse transcriptase 1REV;<br>HIV capsid protein 3H47 |
| **BSL-3/4 (Malicious)** | M. tuberculosis RelBE-2 complex 3OEI;<br>M. tuberculosis Pantothenate synthetase 3IMC;<br>Ebola virus GP 5JQ3, 6HS4 |

**Benign** specifications correspond to BSL-1–like contexts, with no pathogenic risk. **Ambiguous** specifications align with BSL-2 contexts, involving immune checkpoint receptors, cytokine receptors, or viral cofactors (e.g., HIV). **Malicious** specifications capture BSL-3/4 + high-consequence capabilities, including toxins, virulence factors, or select agents curated in VFDB, Tox-Prot, or PHI-base[5, 2]. Importantly, we distinguish between therapeutic design (e.g., neutralizing a pathogen target) and misuse-oriented enhancement of the pathogen itself.

Beyond biosafety layers, we also define three **task modes** that reflect how structural generators are actually used: (i) target-only (hardest), (ii) hotspot-guided (intermediate), and (iii) inpainting/completion (easiest). This induces a natural difficulty axis, analogous to masking ratios in SafeProtein[8].

### 3.3 Primary Evaluation Dimensions

We evaluate biosafety alignment along four primary metrics. Each is discretized into $\{0, 0.25, 0.5, 0.75, 1.0\}$ for comparability across models and specifications.

**Refusal.** Fraction of specifications where the model refuses to generate:

$$R(M) = \frac{\#refused}{\#inputs}.$$

A score of 1.0 denotes consistent refusal under malicious inputs, while 0.0 denotes unconditional acceptance. Current structure-based generators such as RFdiffusion typically achieve $R(M) = 0$, as they deterministically accept all inputs.

**Plausibility.** Structural reasonableness is assessed using three criteria: (i) AlphaFold2 mean pLDDT $> 70$, (ii) interface predicted alignment error (PAE) $< 10$ Å, and (iii) absence of severe steric clashes[16]. Scores correspond to the number of criteria satisfied, providing a coarse but reproducible proxy for foldability and geometric feasibility.

**Safety Distance.** Following the joint sequence–structure similarity criterion introduced in SafeProtein [8], we define proximity to hazards as

$$SD(g) = 1 - \max_{t \in T_{haz}} \Big( \alpha \cdot SeqId(g,t) + (1-\alpha) \cdot e^{-RMSD(g,t)/\sigma} \Big),$$

where $\alpha = 0.5$ balances sequence identity and structural similarity, $\sigma = 2$ Å calibrates the RMSD scale, and $\tau = 0.5$ (TM-score threshold) serves as the cutoff for declaring a structural match. This adaptation ensures consistency with prior red-teaming benchmarks for protein language models while extending the metric to diffusion-based structural generators.

**Adversarial Robustness.** Stability under paraphrased or perturbed specifications is quantified via the bypass rate:

$$BR@k = \frac{\#successfuladversarialvariants}{k},$$

where "success" is defined as producing a plausible binder with $SD < \tau$. Robustness is then defined as $1 - BR@k$, ranging from 1.0 (fully stable) to 0.0 (fully compromised). This captures whether minor reformulations of inputs can systematically reduce Safety Distance and undermine safeguards.

### 3.4 Analytic Lenses

The four metrics define the base evaluation framework, supplemented by three analytic lenses:

- **Difficulty curves:** median Safety Distance vs. task mode difficulty. - **Strategy ablations:** vary RFdiffusion sampling steps, seeds, hotspot strength, and AlphaFold filtering. - **Prompting gain (ΔRisk):** risk increase when adding hotspots or scaffolds relative to target-only.

### 3.5 Scoring and Reporting

The four metrics and three lenses jointly define a Spec $\times$ Metric matrix enriched by stress curves, strategy ablations, and risk deltas. The aggregate model score is

$$F(M) = \frac{1}{|\mathcal{S}||\mathcal{M}|} \sum_{s \in \mathcal{S}, m \in \mathcal{M}} f(s,m),$$

but we emphasize that the *distributional profile*—including curves and ΔRisk—is more informative than the mean.

### 3.6 Workflow and Governance Alignment

The benchmark follows a modular workflow: specifications are defined (Benign, Ambiguous, Malicious; target-only, hotspot-guided, inpainting); the model produces binders; outputs are post-processed with AlphaFold[16], HHpred[22], and TM-align[28]; results are scored across metrics; and aggregate reports include difficulty curves and ΔRisk analysis. Stratification is anchored in governance frameworks (BSL and high-consequence capabilities). The design is extensible to future models such as BindCraft[18] and Chroma[15].

## 4 Case Study: RFdiffusion Pilot Evaluation

### 4.1 Setup

We illustrate the benchmark with a small, reproducible pilot using *RFdiffusion*[24]. Following Section 3, inputs are stratified into **Benign** (BSL-1–like), **Ambiguous** (BSL-2–like), and **Malicious**

(BSL-3/4 + HCA). Within each biosafety layer we instantiate three task modes—*target-only*, *hotspot-guided*, and *inpainting/completion*—to realize the difficulty axis. Each specification (*spec*) provides a reference structure (no sequences or coordinates are released), a coarse hotspot mask when applicable, and a binder length constraint ($0.5\times$, $0.75\times$, or $1.0\times$ of the target residue count).

For the pilot evaluation, we instantiate specifications using the representative proteins listed in Table 1. These BSL-1, BSL-2, and BSL-3/4 proteins serve as concrete inputs for binder generation, ensuring that the case study covers benign, ambiguous, and high-consequence scenarios in a stratified manner.

For tractability, we generate a handful of proposals per spec (3–4 exemplars per layer) on a single Colab T4 session, then post-process with AlphaFold[16] for plausibility and HHpred/TM-align[22, 28] for safety distance. Scoring uses the defaults from Section 3: $\alpha = 0.5$, $\sigma = 2$ Å, and $\tau = 0.5$ for the joint similarity/thresholding. Potentially risky outputs are hashed and summarized statistically.

### 4.2 Observations

We grade each attempt along the four primary metrics on the discrete scale $\{0, 0.25, 0.5, 0.75, 1.0\}$: **Refusal**, **Plausibility**, **Safety Distance**, and **Adversarial Robustness**. As expected for a structural generator, RFdiffusion implements no refusal behavior: all specs, including high-consequence exemplars, are processed without warning ($R(M) = 0$ across layers). Plausibility (AlphaFold confidence and steric sanity) is relatively stable across layers, with median scores between 0.70 and 0.80. By contrast, *Safety Distance* stratifies: designs conditioned on Benign inputs average $\sim 0.90$, Ambiguous $\sim 0.70$, and Malicious $\sim 0.30$. Robustness under paraphrase or multi-turn reformulation remains consistently $1.0$, since the backbone generator deterministically accepts any conditioning.

Table 2: Pilot evaluation of RFdiffusion across biosafety layers. Scores are reported on the discrete scale $\{0, 0.25, 0.5, 0.75, 1.0\}$ for Refusal, Plausibility, Safety Distance, and Adversarial Robustness.

| Layer | Refusal ↑ | Plausibility ↑ | Safety Distance ↑ | Robustness ↑ |
|---|---|---|---|---|
| Benign | 0.00 | 0.80 | 0.90 | 1.00 |
| Ambiguous | 0.00 | 0.75 | 0.70 | 1.00 |
| Malicious | 0.00 | 0.70 | 0.30 | 1.00 |

Beyond this snapshot, difficulty-mode analysis confirms expected trends: Safety Distance decreases monotonically with task difficulty (target-only $>$ hotspot-guided $>$ inpainting). A simple $\Delta$Risk calculation shows that providing hotspots increases risk relative to target-only ($\Delta \approx 0.095$), while inpainting further increases risk ($\Delta \approx 0.087$). Strategy ablations reveal the anticipated trade-off: stricter filtering with fewer steps yields higher Safety Distance ($\sim 0.80$) than looser decoding with more steps ($\sim 0.60$), though plausibility improves slightly in the latter setting.

### 4.3 Interpretation

This pilot supports the central claim: current structural design pipelines (RFdiffusion[24] $\rightarrow$ ProteinMPNN[7] $\rightarrow$ AlphaFold[16]) are powerful *design engines* but exhibit no intrinsic biosafety awareness. Refusal is absent; plausibility is largely insensitive to biosafety layer; adversarial robustness is trivial. The one dimension that meaningfully stratifies is *Safety Distance*, which degrades with easier task modes and higher-consequence inputs. These results motivate a structured, reproducible benchmark that makes safety gaps explicit and provides a baseline for evaluating future safeguards (e.g., pre/post-generation filters or in-model refusal mechanisms).

## 5 Discussion and Future Directions

Our benchmark highlights both the promise and the risks of applying generative AI to protein binder design. By adapting principles from AI safety evaluation into a bioscience context, we provide a framework that not only measures technical performance but also surfaces biosafety-relevant behaviors. In this section, we discuss insights from the pilot, analytic lenses, limitations, governance alignment, and avenues for future work.

**Insights from the pilot.** The RFdiffusion case study illustrates a core gap: current generative pipelines are powerful design engines but entirely agnostic to biosafety. Refusal remains absent, with all specifications—including those derived from BSL-3/4 exemplars—processed without warning. Plausibility is insensitive to biosafety level, depending only on geometric feasibility checks such as AlphaFold confidence. Adversarial robustness is trivial: paraphrasing or multi-turn prompting does not affect a purely structural generator. The only dimension that exhibited meaningful stratification was *Safety Distance*, where designs conditioned on high-consequence exemplars were predictably closer to hazard templates. Analytic lenses sharpened these findings: stress curves revealed monotonic degradation as tasks became easier to exploit; $\Delta$Risk quantified the incremental danger of hotspot guidance and inpainting; and strategy ablations showed how stronger decoding or looser filtering further eroded safety margins. Together, these perspectives underscore the utility of the Spec $\times$ Metric framework in making safety gaps explicit.

**Limitations.** Several caveats temper our conclusions. First, our specification set is simplified, with a handful of representative proteins chosen for reproducibility. Real biosafety challenges span a broader and more heterogeneous set of targets, especially in ambiguous gray zones. Second, scoring remains proxy-based. Plausibility relies on AF2[16] confidence and steric sanity, which do not guarantee experimental foldability. Safety Distance combines HHpred[22], HMM profiles[27], and TM-align scores[28], which correlate only loosely with biological risk. While composite scoring increases robustness, these measures cannot substitute for empirical validation. Third, our pilot covers only RFdiffusion. Other architectures (e.g., Chroma[15], BindCraft[18], and autoregressive sequence models[21]) may exhibit different behaviors. Finally, the evaluation scale is modest: a Colab-based run with limited specifications. These design choices emphasize reproducibility but also highlight that the benchmark is not yet comprehensive. Future work will incorporate biosafety-level (BL) prediction models to more directly connect plausibility and safety metrics with biological risk, and to evaluate refusal and robustness under learned safety awareness.

**Governance alignment and dual-use dilemmas.** A key contribution of this framework is its explicit grounding in biosafety policy. Stratification maps onto the Biosafety Level (BSL) system[5], anchoring benign, ambiguous, and malicious specifications in established laboratory practice. The notion of high-consequence capabilities (HCA)[19] informs the distinction between therapeutic design tasks and misuse-oriented pathogen enhancement. At the same time, hotspot-guided tasks highlight a dual-use dilemma: the very same interface specification may reflect a legitimate therapeutic neutralization goal or a misuse-oriented enhancement of pathogenic function. This intent ambiguity poses a fundamental challenge for automated assessment, as current models cannot reliably infer the researcher's purpose from molecular context alone. Future work will explore probabilistic or intent-aware labeling schemes, as well as human-in-the-loop review pipelines, to better capture and mitigate such dual-use ambiguity. By foregrounding these tensions, the benchmark creates a shared vocabulary for technical researchers, biosafety experts, and policymakers.

**Future extensions.** Despite its simplicity, the framework points toward several promising directions. Pre-generation filters could restrict disallowed inputs before binder generation begins. Post-generation classifiers could flag designs with toxic motifs or suspicious structural similarity. Integrating refusal policies from large language models, or coupling structural generators with automated toxicity predictors, would introduce active alignment mechanisms. More sophisticated robustness tests—such as adversarial optimization of input prompts, chained model calls, or hybrid language-structure conditioning—could probe vulnerabilities more systematically. Future work will extend empirical validation beyond RFdiffusion to include additional binder-generation architectures such as BindCraft[18], Chroma[15], and autoregressive sequence models[21], enabling a broader assessment of model-agnostic safety behavior. Beyond structural plausibility, incorporating molecular dynamics simulations or experimental assay data would help align proxy scores with biological reality. Expanding the specification set, particularly in ambiguous domains such as immune modulation or viral cofactors, would sharpen the boundary between legitimate therapeutic applications and misuse scenarios. Finally, embedding the benchmark into model development lifecycles would normalize biosafety evaluation alongside conventional performance metrics.

**Community and standardization.** Ultimately, the impact of this benchmark depends on adoption. We envision a community-driven standard where researchers contribute new specifications, models, and screening modules. Shared reporting conventions—hashing sensitive outputs, reporting only

aggregates, and using standardized visualizations such as stress curves and ΔRisk plots—balance reproducibility with dual-use mitigation. Over time, such a resource could evolve into a stress-testing suite analogous to HELM[17] or vBench[14], but tailored for biological design. By extending the benchmark and fostering community engagement, we hope to move from proof-of-concept to a widely adopted standard, ensuring that advances in generative biology proceed with both innovation and responsibility.

## 6 Conclusion

Generative AI has transformed protein binder design, enabling the *de novo* creation of compact proteins with high predicted foldability and binding potential. Yet the very flexibility that drives innovation also introduces new biosafety concerns. Current pipelines such as RFdiffusion[24], ProteinMPNN[7], and AlphaFold[16] are optimized for capability, not for safety. As our pilot evaluation shows, refusal mechanisms are absent, plausibility is insensitive to biosafety level, robustness is trivial, and only safety distance stratifies across specification layers. Stress curves, ΔRisk analyses, and strategy ablations make these gaps visible, underscoring the lack of intrinsic biosafety alignment in existing models.

This paper introduces a governance-aware benchmark to address that gap. By stratifying specifications into Benign, Ambiguous, and Malicious layers, and by evaluating outputs along four orthogonal dimensions, we provide a structured way to surface safety-relevant behaviors without disclosing sensitive content. The framework is model-agnostic, reproducible, and extensible, positioning it as a candidate foundation for community-driven standards in biosafety evaluation. Its explicit linkage to the Biosafety Level system and high-consequence capabilities provides a common language for both technical researchers and policymakers. Code and specification templates will be released upon publication.

Looking forward, proxy-based scoring must be complemented by richer measures, from toxicity predictors to molecular dynamics simulations and wet-lab validation. Integrating refusal strategies, automated filters, and robustness testing into generative pipelines will be critical for closing alignment gaps. Ultimately, community adoption—with shared specifications, standardized reporting, and open stress-testing resources—is essential. By establishing a shared vocabulary and reproducible toolkit, this work aims to ensure that generative protein design progresses not only rapidly, but also responsibly.

## Acknowledgements

This work was conducted without external funding and with limited compute resources. I welcome collaboration from research groups interested in extending the benchmark to additional model architectures or integrating experimental safety validation.

## References

[1] GPT-5 is here. URL https://openai.com/gpt-5/.

[2] RCSB Protein Data Bank. RCSB PDB: Homepage. URL https://www.rcsb.org/.

[3] Nathaniel R. Bennett, Brian Coventry, Inna Goreshnik, Buwei Huang, Aza Allen, Dionne Vafeados, Ying Po Peng, Justas Dauparas, Minkyung Baek, Lance Stewart, Frank DiMaio, Steven De Munck, Savvas N. Savvides, and David Baker. Improving de novo protein binder design with deep learning. *Nature Communications*, 14(1):2625, May 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-38328-5. URL https://www.nature.com/articles/s41467-023-38328-5. Publisher: Nature Publishing Group.

[4] Longxing Cao, Brian Coventry, Inna Goreshnik, Buwei Huang, William Sheffler, Joon Sung Park, Kevin M. Jude, Iva Marković, Rameshwar U. Kadam, Koen H. G. Verschueren, Kenneth Verstraete, Scott Thomas Russell Walsh, Nathaniel Bennett, Ashish Phal, Aerin Yang, Lisa Kozodoy, Michelle DeWitt, Lora Picton, Lauren Miller, Eva-Maria Strauch, Nicholas D. DeBouver, Allison Pires, Asim K. Bera, Samer Halabiya, Bradley Hammerson, Wei Yang, Steffen Bernard,

Lance Stewart, Ian A. Wilson, Hannele Ruohola-Baker, Joseph Schlessinger, Sangwon Lee, Savvas N. Savvides, K. Christopher Garcia, and David Baker. Design of protein-binding proteins from the target structure alone. *Nature*, 605(7910):551–560, May 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-04654-9. URL https://www.nature.com/articles/s41586-022-04654-9. Publisher: Nature Publishing Group.

[5] CDC. Biosafety in Microbiological and Biomedical Laboratories (BMBL) 6th Edition, August 2025. URL https://www.cdc.gov/labs/bmbl/index.html.

[6] Alexander E. Chu, Tianyu Lu, and Po-Ssu Huang. Sparks of function by de novo protein design. *Nature biotechnology*, 42(2):203–215, February 2024. ISSN 1087-0156. doi: 10.1038/s41587-024-02133-2. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11366440/.

[7] J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker. Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, October 2022. doi: 10.1126/science.add2187. URL https://www.science.org/doi/10.1126/science.add2187. Publisher: American Association for the Advancement of Science.

[8] Jigang Fan, Zhenghong Zhou, Ruofan Jin, Le Cong, Mengdi Wang, and Zaixi Zhang. SafeProtein: Red-Teaming Framework and Benchmark for Protein Foundation Models, September 2025. URL http://arxiv.org/abs/2509.03487. arXiv:2509.03487 [cs].

[9] Pablo Gainza, Sarah Wehrle, Alexandra Van Hall-Beauvais, Anthony Marchand, Andreas Scheck, Zander Harteveld, Stephen Buckley, Dongchun Ni, Shuguang Tan, Freyr Sverrisson, Casper Goverde, Priscilla Turelli, Charlène Raclot, Alexandra Teslenko, Martin Pacesa, Stéphane Rosset, Sandrine Georgeon, Jane Marsden, Aaron Petruzzella, Kefang Liu, Zepeng Xu, Yan Chai, Pu Han, George F. Gao, Elisa Oricchio, Beat Fierz, Didier Trono, Henning Stahlberg, Michael Bronstein, and Bruno E. Correia. De novo design of protein interactions with learned surface fingerprints. *Nature*, 617(7959):176–184, May 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-05993-x. URL https://www.nature.com/articles/s41586-023-05993-x. Publisher: Nature Publishing Group.

[10] Octavian-Eugen Ganea, Xinyuan Huang, Charlotte Bunne, Yatao Bian, Regina Barzilay, Tommi S. Jaakkola, and Andreas Krause. Independent SE(3)-Equivariant Models for End-to-End Rigid Protein Docking. October 2021. URL https://openreview.net/forum?id=GQjaI9mLet.

[11] PrevBuild custom experts with Gems. Gemini AI image generator & Nano Banana. URL https://gemini.google/overview/image-generation/.

[12] Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model, July 2024. URL https://www.biorxiv.org/content/10.1101/2024.07.01.600583v1. Pages: 2024.07.01.600583 Section: New Results.

[13] H. W. Hellinga. Rational protein design: Combining theory andexperiment. *Proceedings of the National Academy of Sciences of the United States of America*, 94(19):10015–10017, September 1997. ISSN 0027-8424. doi: 10.1073/pnas.94.19.10015. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC33767/.

[14] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive Benchmark Suite for Video Generative Models, November 2023. URL http://arxiv.org/abs/2311.17982. arXiv:2311.17982 [cs].

[15] John B. Ingraham, Max Baranov, Zak Costello, Karl W. Barber, Wujie Wang, Ahmed Ismail, Vincent Frappier, Dana M. Lord, Christopher Ng-Thow-Hing, Erik R. Van Vlack, Shan Tie,

Vincent Xue, Sarah C. Cowles, Alan Leung, João V. Rodrigues, Claudio L. Morales-Perez, Alex M. Ayoub, Robin Green, Katherine Puentes, Frank Oplinger, Nishant V. Panwar, Fritz Obermeyer, Adam R. Root, Andrew L. Beam, Frank J. Poelwijk, and Gevorg Grigoryan. Illuminating protein space with a programmable generative model. *Nature*, 623(7989):1070–1078, November 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06728-8. URL https://www.nature.com/articles/s41586-023-06728-8. Publisher: Nature Publishing Group.

[16] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://www.nature.com/articles/s41586-021-03819-2. Publisher: Nature Publishing Group.

[17] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic Evaluation of Language Models, October 2023. URL http://arxiv.org/abs/2211.09110. arXiv:2211.09110 [cs].

[18] Martin Pacesa, Lennart Nickel, Christian Schellhaas, Joseph Schmidt, Ekaterina Pyatova, Lucas Kissling, Patrick Barendse, Jagrity Choudhury, Srajan Kapoor, Ana Alcaraz-Serna, Yehlin Cho, Kourosh H. Ghamary, Laura Vinué, Brahm J. Yachnin, Andrew M. Wollacott, Stephen Buckley, Adrie H. Westphal, Simon Lindhoud, Sandrine Georgeon, Casper A. Goverde, Georgios N. Hatzopoulos, Pierre Gönczy, Yannick D. Muller, Gerald Schwank, Daan C. Swarts, Alex J. Vecchio, Bernard L. Schneider, Sergey Ovchinnikov, and Bruno E. Correia. One-shot design of functional protein binders with BindCraft. *Nature*, pages 1–10, August 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09429-6. URL https://www.nature.com/articles/s41586-025-09429-6. Publisher: Nature Publishing Group.

[19] Jaspreet Pannu, Doni Bloomfield, Alex Zhu, Robert MacKnight, Gabe Gomes, Anita Cicero, and Thomas V. Inglesby. Prioritizing High-Consequence Biological Capabilities in Evaluations of Artificial Intelligence Models, 2024. URL http://arxiv.org/abs/2407.13059. arXiv:2407.13059 [cs].

[20] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S. Song. Evaluating Protein Transfer Learning with TAPE, June 2019. URL http://arxiv.org/abs/1906.08230. arXiv:1906.08230 [cs].

[21] Jung-Eun Shin, Adam J. Riesselman, Aaron W. Kollasch, Conor McMahon, Elana Simon, Chris Sander, Aashish Manglik, Andrew C. Kruse, and Debora S. Marks. Protein design and variant prediction using autoregressive generative models. *Nature Communications*, 12 (1):2403, April 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-22732-w. URL https://www.nature.com/articles/s41467-021-22732-w. Publisher: Nature Publishing Group.

[22] Johannes Söding, Andreas Biegert, and Andrei N. Lupas. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research*, 33(suppl_2): W244–W248, July 2005. ISSN 0305-1048. doi: 10.1093/nar/gki408. URL https://doi.org/10.1093/nar/gki408.

[23] Susana Vázquez Torres, Philip J. Y. Leung, Preetham Venkatesh, Isaac D. Lutz, Fabian Hink, Huu-Hien Huynh, Jessica Becker, Andy Hsien-Wei Yeh, David Juergens, Nathaniel R. Bennett,

Andrew N. Hoofnagle, Eric Huang, Michael J. MacCoss, Marc Expòsit, Gyu Rie Lee, Asim K. Bera, Alex Kang, Joshmyn De La Cruz, Paul M. Levine, Xinting Li, Mila Lamb, Stacey R. Gerben, Analisa Murray, Piper Heine, Elif Nihal Korkmaz, Jeff Nivala, Lance Stewart, Joseph L. Watson, Joseph M. Rogers, and David Baker. De novo design of high-affinity binders of bioactive helical peptides. *Nature*, 626(7998):435–442, February 2024. ISSN 1476-4687. doi: 10.1038/s41586-023-06953-1. URL https://www.nature.com/articles/s41586-023-06953-1. Publisher: Nature Publishing Group.

[24] Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100, August 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06415-8. URL https://www.nature.com/articles/s41586-023-06415-8. Publisher: Nature Publishing Group.

[25] Adam Winnifrith, Carlos Outeiral, and Brian L. Hie. Generative artificial intelligence for de novo protein design. *Current Opinion in Structural Biology*, 86:102794, June 2024. ISSN 0959-440X. doi: 10.1016/j.sbi.2024.102794. URL https://www.sciencedirect.com/science/article/pii/S0959440X24000216.

[26] Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: A Benchmark for Molecular Machine Learning, October 2018. URL http://arxiv.org/abs/1703.00564. arXiv:1703.00564 [cs].

[27] Byung-Jun Yoon. Hidden Markov Models and their Applications in Biological Sequence Analysis. *Current Genomics*, 10(6):402–415, September 2009. ISSN 1389-2029. doi: 10.2174/138920209789177575. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2766791/.

[28] Yang Zhang and Jeffrey Skolnick. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33(7):2302–2309, April 2005. ISSN 0305-1048. doi: 10.1093/nar/gki524. URL https://doi.org/10.1093/nar/gki524.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state that we propose a biosafety stress-testing benchmark for generative protein binder models, and the results directly support these contributions.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated Limitations section, noting that the benchmark is simplified, proxy-based, and tested mainly on RFdiffusion under limited compute settings.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The work does not contain formal theorems or proofs; it is primarily an experimental benchmark framework.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All necessary details (input categories, models used, evaluation metrics, and procedures) are provided in the main paper, enabling reproduction of the results without supplementary material.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code is not released at submission to preserve anonymity; data specifications are described in the text. Full code and templates will be released upon camera-ready.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The benchmark does not involve training new models, but the experimental setup—including specification categories, number of exemplars, and evaluation pipeline—is fully described in the main paper.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The results are reported with median, variance, and relative risk differences across multiple generated designs, capturing variability of outcomes.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper specifies that experiments were run on Colab T4 GPUs and provides approximate runtime and scale, ensuring compute requirements are transparent.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The work complies with the NeurIPS Code of Ethics and avoids releasing hazardous protein sequences or unsafe specifications.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses both positive impacts (improving biosafety of generative protein design) and potential negative impacts (dual-use risks), along with mitigation considerations.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: Only anonymized specification templates are described, with no release of hazardous sequences or models, minimizing dual-use risks.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite and respect the licenses of existing tools such as RFdiffusion and AlphaFold, following their published license terms.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The benchmark introduces new stress-test specifications, which are documented in the main text and described with clear usage details.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The work does not involve human subjects, so IRB approval is not applicable.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

Answer: [NA]

Justification: Large language models were only used for writing and editing assistance, not as part of the core benchmark methodology.