Structural Memorization in AlphaFold: Adversarial Mutations Reveal Template Reliance, Confidence Failures, and Implications for Protein Design

Anonymous Authors[1]

Abstract

AlphaFold has transformed structural biology and spawned an ecosystem of derivative tools for protein design, binding prediction, and drug discovery. Whether AlphaFold has learned generalizable biophysical principles versus template-based pattern matching remains unclear—a distinction critical for applications beyond its training context. Here, we perform a systematic adversarial evaluation of AlphaFold 3 using point and deletion mutations across 200 proteins, including experimentally validated fold-switching proteins. Predicted structures remain invariant to mutations of up to 40% of residues—including deliberately destabilizing substitutions—and to deletions of 10%, even for fold-switching proteins known to adopt alternative conformations under precisely such perturbations and for which AlphaFold is expected to perform best. Confidence metrics prove unreliable, selecting the most accurate structure at most 35% of the time and correlating with training-set template quality rather than biophysical prediction accuracy—suggesting AlphaFold's uncertainty estimates reflect template availability more than genuine structural reasoning. ESMFold exhibits greater mutational sensitivity. Together, these findings suggest AlphaFold relies heavily on memorized templates rather than biophysical reasoning, with profound implications for the reliability of AlphaFold-based protein design, drug discovery, and modeling workflows.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by *The Impact of Memorization on Trustworthy Foundation Models* Workshop @ ICML. Do not distribute.

1. Introduction

AlphaFold 2 demonstrated that deep learning could predict three-dimensional protein structures from sequence alone with near-experimental accuracy (Jumper et al., 2021; Bertoline et al., 2023; Evans et al., 2022), and AlphaFold 3 extended these capabilities to nucleic acids, small molecules, and complete biomolecular assemblies (Abramson et al., 2024; Feldman & Skolnick, 2025; Thompson & Petrić Howe, 2024). This success has spawned a broad ecosystem of derivative methods for protein design, binder generation, and stability and affinity prediction that incorporate AlphaFold architectures or predictions as core components (Watson et al., 2023; Stark et al., 2025; Pacesa et al., 2024; Liu et al., 2025; Deng et al., 2025; Wee & Wei, 2024; Yang et al., 2023; Dauparas et al., 2022; Wohlwend et al., 2024; Zhang et al., 2024; Brixi et al., 2025; Hu & Ohue, 2025; Borkakoti & Thornton, 2023; Fadini et al., 2025).

Whether AlphaFold has learned true biophysical principles or relies primarily on template matching—similar to a threading algorithm (Mirny & Shakhnovich, 1998)—is therefore a question of broad scientific importance, as derivative methods may inherit the same constraints. Evidence for the latter is mounting: performance degrades on low-similarity inputs (Agarwal & McShan, 2024; Li et al., 2025), fold-switching proteins converge to a single dominant fold (Chakravarty et al., 2024; 2025), and design pipelines generate sequences that appear well-folded in silico but fail experimentally (Molodenskiy et al., 2025; Garcia et al., 2025; Feldman & Feldman, 2025; Dieckhaus & Kuhlman, 2024; Rose et al., 2024; Gut & Lemmin, 2024). Distinguishing these hypotheses requires adversarial evaluation: a model grounded in biophysics should respond to mutations in proportion to their structural impact, whereas a template-driven model may remain invariant despite substantial, nonphysical sequence divergence.

Here, we perform a systematic adversarial evaluation of AlphaFold 3 using complementary point-mutation and deletion-mutation regimes across 200 proteins, including proteins with and without training-set homologs and experimentally validated fold-switching proteins. We compare

AlphaFold 3 with ESMFold (Lin et al., 2023) and evaluate confidence metrics relative to AlphaFold 2. Full methodological details are provided in Section B.

2. Results

2.1. Adversarial Point Mutation Analysis

We conducted an adversarial point mutational analysis across 200 protein sequences spanning four categories—monomer-novel, monomer-similar, multimer-novel, and multimer-similar—where "similar" denotes greater than 30% sequence homology to the AlphaFold 3 training set and "novel" indicates less than 30% homology. Each sequence was mutated at 5, 10, 20, 40, and 70% thresholds using maximally disruptive substitutions biased toward core positions (Yu et al., 2024; Feldman et al., 2025); full details are in Section B. Predicted structures were compared to the unmutated prediction using the template modeling score (TM-score) (Zhang & Skolnick, 2005; 2004) and DockQ (Basu & Wallner, 2016), which measure global and interfacial protein structure alignments, respectively.

AlphaFold 3 maintains global fold similarity up to remarkably high perturbation levels (Figure 1). At 40% mutation, monomeric proteins still preserve the same global fold on average (TM-score ≥ 0.5), and multimeric proteins do so in over 40% of cases—with no discernible difference between novel and similar proteins. TM-scores decline only gradually through 70% mutation, while interfacial geometry degrades far more rapidly, as by 20% mutation, approximately half of interfacial structures are incorrect relative to the unmutated prediction. Confidence metrics are similarly insensitive, with ranking scores remaining above the high-confidence threshold of 0.6 (Evans et al., 2022; Wee & Wei, 2024) through 20% mutation despite deliberate, large-scale sequence disruption (Pearson $r = -0.558$, Spearman $\rho = -0.544$).

Metric	Corr.	Coeff.	5%	10%	20%	40%	70%	Mean
TM-score	Pearson	−0.66	0.77	0.70	0.63	0.49	0.31	0.58
	Spearman	−0.66	—	—	—	—	—	—
Rank. sc.	Pearson	−0.57	0.77	0.74	0.70	0.62	0.51	0.67
	Spearman	−0.55	—	—	—	—	—	—

Table 1. **Structural and confidence metrics for fold-switching proteins under point mutations.** TM-score and AlphaFold 3 ranking score correlations with mutation percentage and mean metric values at each threshold, computed across fifteen experimentally validated fold-switching proteins (Porter & Looger, 2018). See Section A.

We extended this analysis to fifteen experimentally validated monomeric fold-switching proteins (Porter & Looger, 2018), directing mutations at residues empirically shown to induce conformational switching. As summarized in Table 1, predicted structures remained highly invariant through the 40% threshold (mean TM-score ≥ 0.63), indistinguishable in behavior from non-fold-switching proteins. AlphaFold 3 thus fails to capture biologically plausible mutational responses even when targeted at known fold-switching sites in proteins where it is expected to perform best, suggesting it should be used with caution in sequence optimization workflows.

2.2. Adversarial Deletion Mutation Analysis

We conducted an analogous experiment using residue deletions, which are typically far more disruptive than point mutations—even sparse deletions can cause severe destabilization or collapse of tertiary structure (Shah et al., 2015; Huss et al., 2021). Accordingly, lower thresholds of 1, 3, 5, and 10% were employed, with deletions again biased toward central positions. Full details are in Section B.

Despite the severity of this perturbation type, AlphaFold 3 again maintains global fold integrity (Figure 2). At 10% deletions, an overwhelming majority of monomeric proteins retain accurate predicted structures across all tested thresholds, while multimeric proteins degrade more rapidly, crossing into low-confidence regimes by 5%—yet again with no discernible difference between novel and similar proteins. Confidence metrics are more responsive than in the point-mutation regime but remain elevated at low deletion levels: mean ranking score declines from 0.67 at 1% to 0.60 at 3% and 0.50 at 5% (Pearson $r = -0.514$, Spearman $\rho = -0.491$), despite sparse core deletions being expected to substantially impair structural integrity in real proteins (Woods et al., 2023; Guan et al., 2025).

2.3. Comparison of AlphaFold 3 and ESMFold Under Adversarial Mutation

To determine whether structural invariance is specific to AlphaFold 3 or reflects a broader property of deep learning–based structure prediction, we compared both models across the same mutational regimes, restricting to monomeric proteins as ESMFold does not support multimeric inputs. Full details are in Section B.

The two models behave comparably under deletion mutations, with ESMFold exhibiting slightly greater structural flexibility across all thresholds (Figure 3). The contrast is far more pronounced for point mutations: ESMFold's accuracy collapses rapidly between 20% and 40% mutation, while AlphaFold 3 maintains a gradual decline until 70%— corresponding to only 30% sequence identity—at which point both models converge. This pattern holds for the fifteen fold-switching proteins, where ESMFold again diverges substantially beyond 10% mutation, though less pronouncedly than in the full dataset (Appendices; Table 2). ESMFold's earlier structural response to sequence perturbation suggests it better captures sequence–structure coupling. Conversely,

AlphaFold 3's persistence may reflect a stronger reliance on learned templates that override sequence-level signals even as the input diverges substantially from biophysically viable configurations.

2.4. Evaluation of Confidence Metrics in AlphaFold 2 versus AlphaFold 3

We evaluated both models on 100 novel proteins—50 monomers and 50 multimers—comparing predicted structures against experimental PDB structures to assess confidence metric reliability (Abramson et al., 2024; Jumper et al., 2021; Feldman & Skolnick, 2025). AlphaFold 3 achieves significantly higher structural accuracy than AlphaFold 2, particularly for multimeric complexes (Feldman & Skolnick, 2025; Abramson et al., 2024), but both models fail to reliably identify their most accurate prediction (Figure 4): AlphaFold 3 selects the best structure by TM-score or DockQ only ∼25% of the time, AlphaFold 2 at most 35%. Incorrect selection is not inconsequential, with accuracy losses reaching up to 0.17 in TM-score and 0.40 in DockQ in the worst cases, even if averages are modest (0.01–0.02 TM-score, ∼0.03 DockQ). Ranking score correlates more strongly with accuracy for multimers (Pearson $r = 0.72$ for AlphaFold 3, 0.83 for AlphaFold 2) than monomers ($r = 0.59$ for AlphaFold 3, 0.44 for AlphaFold 2), with AlphaFold 2 outperforming AlphaFold 3 on multimer confidence calibration and vice versa for monomers.

2.5. Structural Template Availability Predicts AlphaFold 3 Confidence

To probe what AlphaFold 3's confidence metrics truly measure, we performed an exhaustive structural search of the pre-cutoff PDB using FoldSeek (van Kempen et al., 2023) and quantified the relationship between template availability and model confidence (Section B). AlphaFold 3's ranking score correlates significantly with the TM-score of the best pre-cutoff structural template (Pearson $r = 0.39$, $p = 7.8 \times 10^{-5}$; Spearman $\rho = 0.39$, $p = 8.2 \times 10^{-5}$), while sequence identity to that template is a non-significant predictor ($r = 0.033$, $p = 0.744$), indicating that confidence tracks structural rather than sequence similarity to training exemplars. This effect is robust across novelty strata for monomers (novel: $r = 0.51$; similar: $r = 0.49$) but absent for similar multimers ($r = 0.14$, $p = 0.34$), suggesting sequence memorization supplants template dependence when the query falls within the training distribution. For monomers, the single best template drives confidence most strongly. For multimers, correlations increase monotonically to $\rho = 0.46$ at the top 100 hits, indicating that breadth of structural coverage matters more than any single exemplar. Full stratified statistics are reported in Table 5.

3. Discussion

The point and deletion mutation analyses presented here reveal a striking invariance in both the structures predicted by AlphaFold 3 and its confidence metrics under substantial sequence perturbations. Up to 40% of residues can be mutated—and up to 10% deleted—before predicted structures meaningfully diverge from the original AlphaFold 3 prediction on average (Zhang & Skolnick, 2004), a trend that persists even for fold-switching proteins experimentally known to adopt multiple conformations (Porter & Looger, 2018). The parsimonious explanation is structural memorization: AlphaFold 3 reproduces a memorized template rather than computing a structure from the input sequence.

The degree of invariance we document suggests that AlphaFold 3 may not be processing input sequences at the granularity required to capture the structural consequences of individual residue changes. Insensitivity to one or two substitutions might reflect genuine fold robustness; insensitivity to 70% adversarial mutation—reducing sequence identity to near zero—is more naturally explained by a model anchored to a memorized exemplar that cannot be dislodged by sequence-level signals alone (Dieckhaus & Kuhlman, 2024).

The structural template analysis provides the most direct quantitative evidence for this memorization hypothesis. AlphaFold 3's confidence correlates significantly with the structural quality of the best pre-cutoff template, while sequence identity to that template is non-significant—indicating that confidence tracks template retrieval quality rather than prediction reliability (Chakravarty et al., 2024). Protein design pipelines that treat AlphaFold confidence as a proxy for designability may therefore be optimizing against memorized templates rather than genuine sequence–structure relationships.

The comparison with ESMFold illuminates the memorization hypothesis further. ESMFold, trained with a masked language modeling objective that encodes coevolutionary constraints without explicit structural templates, exhibits substantially greater mutational sensitivity (Lin et al., 2022). Its earlier structural response demonstrates that sequence–structure coupling is learnable without template retrieval—and that AlphaFold 3's relative insensitivity is a property of its architecture or training regime rather than an inherent ceiling on what deep learning can capture.

These findings also underscore concerns about confidence metric reliability independent of memorization. Neither AlphaFold 2 nor AlphaFold 3 reliably identifies the most accurate structure among five generated predictions, with both models selecting the best structure only approximately 16–35% of the time. Losses from incorrect selection reach up to 0.17 in TM-score and over 0.4 in DockQ—errors that

in drug discovery, protein engineering, or pathway analysis could lead to the exclusion of promising candidates or the retention of problematic ones (Watson et al., 2023; Stark et al., 2025).

These observations do not diminish the transformative impact of AlphaFold on structural biology. The limitations we document are most relevant for out-of-distribution inputs and applications requiring fine-grained sensitivity to sequence variation—precisely the settings in which memorization offers no reliable generalization. The source of this memorization likely combines architectural and data factors: transformer and diffusion frameworks may inherently favor interpolation within the training distribution, and structural databases consist almost exclusively of stable wild-type proteins, providing no signal for mutation-induced structural disruption. Addressing these limitations will require training on destabilized mutants and misfolded proteins, incorporating explicit biophysical priors, and developing confidence estimation methods that decouple uncertainty from template availability.

Impact Statement

This work demonstrates that AlphaFold 3—the dominant foundation model in structural biology—exhibits structural memorization with direct consequences for trustworthy deployment. Its predictions are anchored to training-set templates rather than computed from input sequences, its confidence metrics reflect template availability rather than prediction reliability, and both failure modes propagate into the broad ecosystem of derivative tools that treat AlphaFold outputs as ground truth. A model that memorizes rather than generalizes will appear accurate on standard benchmarks while silently failing on the out-of-distribution inputs that matter most: engineered sequences, destabilizing mutations, and proteins in alternative conformational states. Surfacing these failure modes and connecting them mechanistically to memorization is a necessary step toward structural biology foundation models that are genuinely trustworthy and generalizable.

References

Rcsb protein data bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy | nucleic acids research | oxford academic. URL https://academic.oup.com/nar/article/47/D1/D464/5144139.

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., Beattie, C., Bertolli, O., Bridgland, A., Cherepanov, A., Congreve, M., Cowen-Rivers, A. I., Cowie, A., Figurnov, M., Fuchs, F. B., Gladman, H., Jain, R., Khan, Y. A., Low, C. M. R., Perlin, K., Potapenko, A., Savy, P., Singh, S., Stecula, A., Thillaisundaram, A., Tong, C., Yakneen, S., Zhong, E. D., Zielinski, M., Žídek, A., Bapst, V., Kohli, P., Jaderberg, M., Hassabis, D., and Jumper, J. M. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, June 2024. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-024-07487-w. URL https://www.nature.com/articles/s41586-024-07487-w. Publisher: Nature Publishing Group.

Adrover, M., Pauwels, K., Prigent, S., de Chiara, C., Xu, Z., Chapuis, C., Pastore, A., and Rezaei, H. Prion fibrillization is mediated by a native structural element that comprises helices h2 and h3. *Journal of Biological Chemistry*, 285(27):21004–21012, 2010.

Agarwal, V. and McShan, A. C. The power and pitfalls of alphafold2 for structure prediction beyond rigid globular proteins. *Nature Chemical Biology*, Jun 2024. doi: https://doi.org/10.1038/s41589-024-01638-w. URL https://www.nature.com/articles/s41589-024-01638-w.

Ayed, S. H., Cloutier, A. D., McLeod, L. J., Foo, A. C., Damry, A. M., and Goto, N. K. Dissecting the role of conformational change and membrane binding by the bacterial cell division regulator mine in the stimulation of mind atpase activity. *Journal of Biological Chemistry*, 292(50):20732–20743, 2017.

Basu, S. and Wallner, B. Dockq: A quality measure for protein-protein docking models. *PLOS ONE*, 11(8):e0161879, August 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.0161879. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0161879. Publisher: Public Library of Science.

Bertoline, L. M. F., Lima, A. N., Krieger, J. E., and Teixeira, S. K. Before and after alphafold2: An overview of protein structure prediction. *Frontiers in Bioinformatics*, 3:1120370, February 2023. ISSN 2673-7647. doi: 10.3389/fbinf.2023.1120370. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10011655/.

Biasini, M., Schmidt, T., Bienert, S., Mariani, V., Studer, G., Haas, J., Johner, N., Schenk, A. D., Philippsen, A., and Schwede, T. Openstructure: an integrated software framework for computational structural biology. *Acta Crystallographica Section D: Biological Crystallography*, 69(5):701–709, May 2013. ISSN 0907-4449. doi: 10.

1107/S0907444913007051. URL //journals.iucr.org/paper?ic5090. Publisher: International Union of Crystallography.

Borkakoti, N. and Thornton, J. M. Alphafold2 protein structure prediction: Implications for drug discovery. *Current Opinion in Structural Biology*, 78:102526, 2023. ISSN 0959-440X. doi: https://doi.org/10.1016/j.sbi.2022.102526. URL https://www.sciencedirect.com/science/article/pii/S0959440X22002056.

Brixi, G., Durrant, M. G., Ku, J., Poli, M., Brockman, G., Chang, D., Gonzalez, G. A., King, S. H., Li, D. B., Merchant, A. T., Naghipourfar, M., Nguyen, E., Ricci-Tam, C., Romero, D. W., Sun, G., Taghibakshi, A., Vorontsov, A., Yang, B., Deng, M., Gorton, L., Nguyen, N., Wang, N. K., Adams, E., Baccus, S. A., Dillmann, S., Ermon, S., Guo, D., Ilango, R., Janik, K., Lu, A. X., Mehta, R., Mofrad, M. R. K., Ng, M. Y., Pannu, J., Ré, C., Schmok, J. C., John, J. S., Sullivan, J., Zhu, K., Zynda, G., Balsam, D., Collison, P., Costa, A. B., Hernandez-Boussard, T., Ho, E., Liu, M.-Y., McGrath, T., Powell, K., Burke, D. P., Goodarzi, H., Hsu, P. D., and Hie, B. L. Genome modeling and design across all domains of life with evo 2, February 2025. URL https://www.biorxiv.org/content/10.1101/2025.02.18.638918v1. Pages: 2025.02.18.638918 Section: New Results.

Chakravarty, D., Schafer, J. W., Chen, E. A., Thole, J. F., Ronish, L. A., Lee, M., and Porter, L. L. Alphafold predictions of fold-switched conformations are driven by structure memorization. *Nature Communications*, 15(1), Aug 2024. doi: https://doi.org/10.1038/s41467-024-51801-z. URL https://www.nature.com/articles/s41467-024-51801-z.

Chakravarty, D., Lee, M., and Porter, L. L. Proteins with alternative folds reveal blind spots in alphafold-based protein structure prediction. *Current Opinion in Structural Biology*, 90:102973, 2025. ISSN 0959-440X. doi: https://doi.org/10.1016/j.sbi.2024.102973. URL https://www.sciencedirect.com/science/article/pii/S0959440X24002008.

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Courbet, A., de Haas, R. J., Bethel, N., Leung, P. J. Y., Huddy, T. F., Pellock, S., Tischer, D., Chan, F., Koepnick, B., Nguyen, H., Kang, A., Sankaran, B., Bera, A. K., King, N. P., and Baker, D. Robust deep learning based protein sequence design using proteinmpnn. *bioRxiv*, 2022. doi: 10.1101/2022.06.03.494563. URL https://www.biorxiv.org/content/early/2022/06/04/2022.06.03.494563.

Deng, A., Householder, K., Wu, F., Thrun, S., Garcia, K. C., and Trippe, B. Predicting mutational effects on protein binding from folding energy, July 2025. URL http://arxiv.org/abs/2507.05502. arXiv:2507.05502 [q-bio].

Dieckhaus, H. and Kuhlman, B. Protein stability models fail to capture epistatic interactions of double point mutations. *bioRxiv*, pp. 2024.08.20.608844, August 2024. ISSN 2692-8205. doi: 10.1101/2024.08.20.608844. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11370451/.

Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., Kohli, P., Jumper, J., and Hassabis, D. Protein complex prediction with alphafold-multimer, March 2022. URL https://www.biorxiv.org/content/10.1101/2021.10.04.463034v2. Pages: 2021.10.04.463034 Section: New Results.

Fadini, A., Li, M., McCoy, A. J., Terwilliger, T. C., Read, R. J., Hekstra, D., and AlQuraishi, M. Alphafold as a prior: Experimental structure determination conditioned on a pretrained neural network. *bioRxiv*, 2025. doi: 10.1101/2025.02.18.638828. URL https://www.biorxiv.org/content/early/2025/03/11/2025.02.18.638828.

Feldman, J. and Feldman, T. Resilient biosecurity in the era of ai-enabled bioweapons, August 2025. URL http://arxiv.org/abs/2509.02610. arXiv:2509.02610 [q-bio].

Feldman, J. and Skolnick, J. Af3complex yields improved structural predictions of protein complexes, March 2025. ISSN 1367-4811. URL https://www.biorxiv.org/content/10.1101/2025.02.27.640585v1. Pages: 2025.02.27.640585 Section: New Results.

Feldman, J., Maechler, A., Wang, D., and Shakhnovich, E. Biophysically grounded deep learning improves protein–protein g prediction. *bioRxiv*, 2025. doi: 10.64898/2025.12.23.696257. URL https://www.biorxiv.org/content/early/2025/12/25/2025.12.23.696257.

Gammons, M. V., Renko, M., Johnson, C. M., Rutherford, T. J., and Bienz, M. Wnt signalosome assembly by dep domain swapping of dishevelled. *Molecular cell*, 64(1): 92–104, 2016.

Garcia, M., Dixit, S. M., and Rocklin, G. J. Evaluating zero-shot prediction of protein design

success by alphafold, esmfold, and proteinmpnn. *bioRxiv*, 2025. doi: 10.1101/2025.07.29.667290. URL https://www.biorxiv.org/content/early/2025/08/09/2025.07.29.667290.

Guan, C., Wan, F., Torres, M. D. T., and Fuente-Nunez, C. d. l. Improving functional protein generation via foundation model-derived latent space likelihood optimization, January 2025. URL https://www.biorxiv.org/content/10.1101/2025.01.07.631724v1. Pages: 2025.01.07.631724 Section: New Results.

Gut, J. A. and Lemmin, T. Dissecting alphafold's capabilities with limited sequence information. *bioRxiv*, 2024. doi: 10.1101/2024.03.14.585076. URL https://www.biorxiv.org/content/early/2024/06/25/2024.03.14.585076.

Hamiaux, C., Maddumage, R., Middleditch, M. J., Prakash, R., Brummell, D. A., Baker, E. N., and Atkinson, R. G. Crystal structure of kiwellin, a major cell-wall protein from kiwifruit. *Journal of structural biology*, 187(3): 276–281, 2014.

Hu, W. and Ohue, M. Spatialppiv2: Enhancing protein–protein interaction prediction through graph neural networks with protein language models. *Computational and Structural Biotechnology Journal*, 27:508–518, January 2025. ISSN 2001-0370. doi: 10.1016/j.csbj.2025.01.022. URL https://www.sciencedirect.com/science/article/pii/S2001037025000224.

Huss, P., Meger, A., Leander, M., Nishikawa, K., and Raman, S. Mapping the functional landscape of the receptor binding domain of t7 bacteriophage by deep mutational scanning. *eLife*, 10:e63775, March 2021. ISSN 2050-084X. doi: 10.7554/eLife.63775. URL https://doi.org/10.7554/eLife.63775. Publisher: eLife Sciences Publications, Ltd.

Jones, J., Jiang, W., Synovic, N., Thiruvathukal, G. K., and Davis, J. C. What do we know about hugging face? a systematic literature review and quantitative validation of qualitative claims, 2024. URL https://arxiv.org/abs/2406.08205.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, August 2021. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://www.nature.com/articles/s41586-021-03819-2.

Kadavath, H., Jaremko, M., Jaremko, Ł., Biernat, J., Mandelkow, E., and Zweckstetter, M. Folding of the tau protein on microtubules. *Angewandte Chemie International Edition*, 54(35):10347–10351, 2015.

Li, M. Q. C., Wang, S., Lin, S.-R., Ting, L. E. N., Wan, Z.-H., Xie, G., and Zhang, J. Advantages and limitations of alphafold in structural biology: Insights from recent studies. *The Protein Journal*, Dec 2025. doi: https://doi.org/10.1007/s10930-025-10310-8.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Santos Costa, A. d., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022. doi: 10.1101/2022.07.20.500902. URL https://www.biorxiv.org/content/early/2022/07/21/2022.07.20.500902.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/abs/10.1126/science.ade2574.

Liu, D., Young, F., Lamb, K. D., Claudio Quiros, A., Pancheva, A., Miller, C. J., Macdonald, C., Robertson, D. L., and Yuan, K. Plm-interact: extending protein language models to predict protein-protein interactions. *Nature Communications*, 16(1):9012, October 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-64512-w. URL https://www.nature.com/articles/s41467-025-64512-w. Publisher: Nature Publishing Group.

Lukoyanova, N., Kondos, S. C., Farabella, I., Law, R. H., Reboul, C. F., Caradoc-Davies, T. T., Spicer, B. A., Kleifeld, O., Traore, D. A., Ekkel, S. M., et al. Conformational changes during pore formation by the perforin-related protein pleurotolysin. *PLoS biology*, 13(2):e1002049, 2015.

Mirabello, C. and Wallner, B. Dockq v2: improved automatic quality measure for protein multimers, nucleic acids, and small molecules. *Bioinformatics*, 40 (10):btae586, October 2024. ISSN 1367-4811. doi: 10.1093/bioinformatics/btae586. URL https://doi.org/10.1093/bioinformatics/btae586.

Mirny, L. A. and Shakhnovich, E. I. Protein structure prediction by threading. why it works and

why it does not11edited by f. cohen. *Journal of Molecular Biology*, 283(2):507–526, 1998. ISSN 0022-2836. doi: https://doi.org/10.1006/jmbi.1998.2092. URL https://www.sciencedirect.com/science/article/pii/S0022283698920927.

Molodenskiy, D., Maurer, V. J., Yu, D., Chojnowski, G., Bienert, S., Tauriello, G., Gilep, K., Schwede, T., and Kosinski, J. Alphapulldown2–a general pipeline for high-throughput structural modeling. *Bioinformatics*, 41(3):btaf115, March 2025. ISSN 1367-4811. doi: 10.1093/bioinformatics/btaf115. URL https://doi.org/10.1093/bioinformatics/btaf115.

Morris, V. K., Kwan, A. H., and Sunde, M. Analysis of the structure and conformational states of dewa gives insight into the assembly of the fungal hydrophobins. *Journal of molecular biology*, 425(2):244–256, 2013.

Pacesa, M., Nickel, L., Schmidt, J., Pyatova, E., Schellhaas, C., Kissling, L., Alcaraz-Serna, A., Cho, Y., Ghamary, K. H., Vinué, L., Yachnin, B. J., Wollacott, A. M., Buckley, S., Georgeon, S., Goverde, C. A., Hatzopoulos, G. N., Gönczy, P., Muller, Y. D., Schwank, G., Ovchinnikov, S., and Correia, B. E. Bindcraft: one-shot design of functional protein binders. *bioRxiv*, 2024. doi: 10.1101/2024.09.30.615802. URL https://www.biorxiv.org/content/early/2024/10/01/2024.09.30.615802.

Park, K.-T., Wu, W., Battaile, K. P., Lovell, S., Holyoak, T., and Lutkenhaus, J. The min oscillator uses mind-dependent conformational changes in mine to spatially regulate cytokinesis. *Cell*, 146(3):396–407, 2011.

Pham, C. L., Rey, A., Lo, V., Soulès, M., Ren, Q., Meisl, G., Knowles, T. P., Kwan, A. H., and Sunde, M. Self-assembly of mpg1, a hydrophobin protein from the rice blast fungus that forms functional amyloid coatings, occurs by a surface-driven mechanism. *Scientific reports*, 6 (1):25288, 2016.

Porter, L. L. and Looger, L. L. Extant fold-switching proteins are widespread. *Proceedings of the National Academy of Sciences*, 115(23):5968–5973, June 2018. doi: 10.1073/pnas.1800168115. URL https://www.pnas.org/doi/full/10.1073/pnas.1800168115. Publisher: Proceedings of the National Academy of Sciences.

Poyraz, Ö., Schmidt, H., Seidel, K., Delissen, F., Ader, C., Tenenboim, H., Goosmann, C., Laube, B., Thünemann, A. F., Zychlinsky, A., et al. Protein refolding is required for assembly of the type three secretion needle. *Nature structural & molecular biology*, 17(7):788–792, 2010.

Rao, J. N., Warren, G. Z., Estolt-Povedano, S., Zammit, V. A., and Ulmer, T. S. An environment-dependent structural switch underlies the regulation of carnitine palmitoyltransferase 1a*. *Journal of Biological Chemistry*, 286 (49):42545–42554, 2011a.

Rao, J. N., Warren, G. Z., Estolt-Povedano, S., Zammit, V. A., and Ulmer, T. S. An environment-dependent structural switch underlies the regulation of carnitine palmitoyltransferase 1a*. *Journal of Biological Chemistry*, 286 (49):42545–42554, 2011b.

Rose, T., Zhou, C., and Monti, N. Affinitylm: Binding-site informed multitask language model for drug-target affinity prediction. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 727–734, December 2024. doi: 10.1109/BIBM62325.2024.10822722. URL https://ieeexplore.ieee.org/abstract/document/10822722. ISSN: 2156-1133.

Shah, P., McCandlish, D. M., and Plotkin, J. B. Contingency and entrenchment in protein evolution under purifying selection. *Proceedings of the National Academy of Sciences*, 112(25), June 2015. ISSN 1091-6490. doi: 10.1073/pnas.1412933112. URL http://dx.doi.org/10.1073/pnas.1412933112.

Stark, H., Faltings, F., Choi, M., Xie, Y., Hur, E., O'Donnell, T., Bushuiev, A., Uçar, T., Passaro, S., Mao, W., Reveiz, M., Bushuiev, R., Pluskal, T., Sivic, J., Kreis, K., Vahdat, A., Ray, S., Goldstein, J. T., Savinov, A., Hambalek, J. A., Gupta, A., Taquiri-Diaz, D. A., Zhang, Y., Hatstat, A. K., Arada, A., Kim, N. H., Tackie-Yarboi, E., Boselli, D., Schnaider, L., Liu, C. C., Li, G.-W., Hnisz, D., Sabatini, D. M., DeGrado, W. F., Wohlwend, J., Corso, G., Barzilay, R., and Jaakkola, T. Boltzgen: Toward universal binder design. *bioRxiv*, 2025. doi: 10.1101/2025.11.20.689494. URL https://www.biorxiv.org/content/early/2025/11/24/2025.11.20.689494.

Sweeting, B., Brown, E., Khan, M. Q., Chakrabartty, A., and Pai, E. F. N-terminal helix-cap in α-helix 2 modulates β-state misfolding in rabbit and hamster prion proteins. *PloS one*, 8(5):e63047, 2013.

Szymańska, A., Jankowska, E., Orlikowska, M., Behrendt, I., Czaplewska, P., and Rodziewicz-Motowidło, S. Influence of point mutations on the stability, dimerization, and oligomerization of human cystatin c and its l68q variant. *Frontiers in Molecular Neuroscience*, 5:82, 2012.

Thompson, B. and Petrić Howe, N. Alphafold 3.0: the ai protein predictor gets an upgrade. *Nature*, May 2024. doi: 10.1038/d41586-024-01385-x. URL https://www.nature.com/articles/

d41586-024-01385-x. Bandiera_abtest: a Cg_type: Nature Podcast Publisher: Nature Publishing Group Subject_term: Exoplanets, Physics, Proteomics, Materials science.

van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M., Söding, J., and Steinegger, M. Fast and accurate protein structure search with FoldSeek. *Nature Biotechnology*, May 2023. URL https://www.nature.com/articles/s41587-023-01773-0.

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Hanikel, N., Pellock, S. J., Courbet, A., Sheffler, W., Wang, J., Venkatesh, P., Sappington, I., Torres, S. V., Lauko, A., De Bortoli, V., Mathieu, E., Ovchinnikov, S., Barzilay, R., Jaakkola, T. S., DiMaio, F., Baek, M., and Baker, D. 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.

Wee, J. and Wei, G.-W. Benchmarking alphafold3's protein-protein complex accuracy and machine learning prediction reliability for binding free energy changes upon mutation. *ArXiv*, pp. arXiv:2406.03979v1, June 2024. ISSN 2331-8422. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11177964/.

Wei, Y., Zhang, H., Gao, Z.-Q., Wang, W.-J., Shtykova, E. V., Xu, J.-H., Liu, Q.-S., and Dong, Y.-H. Crystal and solution structures of methyltransferase rsmh provide basis for methylation of c1402 in 16s rrna. *Journal of structural biology*, 179(1):29–40, 2012.

Wohlwend, J., Corso, G., Passaro, S., Reveiz, M., Leidal, K., Swiderski, W., Portnoi, T., Chinn, I., Silterra, J., Jaakkola, T., and Barzilay, R. Boltz-1 democratizing biomolecular interaction modeling, November 2024. URL https://www.biorxiv.org/content/10.1101/2024.11.19.624167v1. Pages: 2024.11.19.624167 Section: New Results.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. Huggingface's transformers: State-of-the-art natural language processing, 2020. URL https://arxiv.org/abs/1910.03771.

Woods, H., Schiano, D. L., Aguirre, J. I., Ledwitch, K. V., McDonald, E. F., Voehler, M., Meiler, J., and Schoeder, C. T. Computational modeling and prediction of deletion mutants. *Structure*, 31(6):713–723.e3, 2023. ISSN 0969-2126. doi: https://doi.org/10.1016/j.str.2023.04.005. URL https://www.sciencedirect.com/science/article/pii/S0969212623001284.

Yang, Z., Zeng, X., Zhao, Y., and Chen, R. Alphafold2 and its applications in the fields of biology and medicine. *Signal Transduction and Targeted Therapy*, 8 (1):1–14, March 2023. ISSN 2059-3635. doi: 10.1038/s41392-023-01381-z. URL https://www.nature.com/articles/s41392-023-01381-z. Publisher: Nature Publishing Group.

Yu, G., Zhao, Q., Bi, X., and Wang, J. Ddaffinity: predicting the changes in binding affinity of multiple point mutations using protein 3d structure. *Bioinformatics*, 40(Supplement_1):i418–i427, July 2024. ISSN 1367-4811. doi: 10.1093/bioinformatics/btae232. URL https://doi.org/10.1093/bioinformatics/btae232.

Zhang, X., Patel, A., Celma, C. C., Yu, X., Roy, P., and Zhou, Z. H. Atomic model of a nonenveloped virus reveals ph sensors for a coordinated process of cell entry. *Nature structural & molecular biology*, 23(1):74–80, 2016.

Zhang, Y. and Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004. ISSN 1097-0134. doi: 10.1002/prot.20264. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.20264. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.20264.

Zhang, Y. and Skolnick, J. Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic Acids Research*, 33(7):2302–2309, 2005. ISSN 0305-1048. doi: 10.1093/nar/gki524. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1084323/.

Zhang, Z., Wayment-Steele, H. K., Brixi, G., Wang, H., Kern, D., and Ovchinnikov, S. Protein language models learn evolutionary statistics of interacting sequence motifs. *Proceedings of the National Academy of Sciences*, 121(45):e2406285121, November 2024. ISSN 1091-6490. doi: 10.1073/pnas.2406285121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2406285121. Publisher: Proceedings of the National Academy of Sciences.

Zhao, K., Chai, X., and Marmorstein, R. Structure and substrate binding properties of cobb, a sir2 homolog protein deacetylase from escherichia coli. *Journal of molecular biology*, 337(3):731–741, 2004.

A. Figures

B. Methods

B.1. AlphaFold Testing Dataset Collection

To rigorously evaluate AlphaFold's performance, we assembled a dataset of 200 proteins, divided evenly into four categories: multimer-novel, multimer-similar, monomer-novel, and monomer-similar, with 50 proteins per category. "Multimeric" and "monomeric" indicate whether a protein contains multiple chains or a single chain, respectively, while "novel" and "similar" denote sequence similarity to proteins in the PDB released on or before September 30, 2021 (noa; Abramson et al., 2024). Proteins labeled as similar share 30% or more sequence homology with pre-September 2021 PDB entries, meaning their homologs could plausibly have been in AlphaFold 3's training set (Feldman & Skolnick, 2025). In contrast, novel proteins have no such homology. Importantly, all proteins in the dataset were released after September 30, 2021, and no two proteins share 30% or more sequence similarity, ensuring minimal redundancy and that none were present in AlphaFold 3's training set. This categorization enables a detailed assessment of AlphaFold's performance across different protein types and levels of novelty.

B.2. Fold-Switching Protein Dataset

To extend the mutational analysis to proteins with experimentally validated alternative conformations, we selected fifteen fold-switching proteins from the curated dataset compiled by Porter et al. (Porter & Looger, 2018). All proteins in this set are monomeric and have been experimentally confirmed to adopt multiple distinct folds under different conditions. The selected proteins correspond to the following PDB entries: 2KXO, 2LSH, 4OV8, 2MZ7, 4PMK, 2N4O, 2KTM, 2LE3, 2X9C, 3J9E, 5SUZ, 4HLS, 1S5P, 3TKA, and 3GAX. For these proteins, only chain A from their corresponding PDB files was used, as described in (Porter & Looger, 2018). Point mutations were introduced to these proteins using procedures similar to those applied to the 200-protein dataset described above, with the exception of a modified position-weighting scheme that prioritized residues or regions known to induce fold-switching. Details of this scheme are provided in Table 4. MSA processing and AlphaFold 3 inference followed the same protocols, with templates excluded. This analysis allowed us to assess whether proteins known to exhibit conformational plasticity show greater structural responsiveness to mutations in AlphaFold 3's predictions compared to proteins without documented fold-switching behavior.

B.3. Mutation Strategy

The mutation percentages applied differed between point mutations and deletion mutations to reflect their relative biophysical severity. Point mutations were tested at 5%, 10%, 20%, 40%, and 70% thresholds, while deletion mutations were tested at 1%, 3%, 5%, and 10% thresholds. This asymmetry reflects the fact that deletion mutations are substantially more deleterious than point mutations (Shah et al., 2015). Even a single residue deletion in a critical region can cause severe structural destabilization or complete loss of fold, whereas point mutations—particularly conservative substitutions—are often tolerated without catastrophic consequences (Huss et al., 2021). By using lower deletion percentages, we ensured that the perturbations remained within a regime where structural prediction could be meaningfully evaluated, while still probing the model's sensitivity to this severe class of mutation.

Mutations were applied cumulatively across all thresholds. Each higher mutation percentage retained all mutations from the previous level and added new mutations to reach the target threshold. For example, the 20% point mutation configuration contained all mutations present in the 10% configuration, plus additional mutations to reach 20% of the sequence length. Similarly, the 5% deletion configuration included all deletions from the 3% level, supplemented with further deletions. This cumulative strategy ensures that differences in structural predictions across mutation levels reflect the incremental addition of perturbations rather than independent sampling at each threshold, enabling a more controlled assessment of mutational sensitivity.

For homomeric complexes, mutations were applied synchronously: the same residue positions were mutated to the same amino acids across all identical chains. For example, in a homodimeric protein, if position 42 was mutated from serine to tryptophan in chain A, position 42 was also mutated from serine to tryptophan in chain B. This approach reflects the biological reality that homomeric complexes consist of identical gene products and thus share identical sequences. In contrast, heteromeric complexes have distinct chains with independent sequences, and mutations were applied independently to each chain, such that different positions and amino acid substitutions were selected for each unique chain type.

B.4. AlphaFold Model Configurations and Inputs

For all mutational analyses, AlphaFold 3 was run using its default configuration, with ten recycling steps and five diffusion samples (Abramson et al., 2024). The primary modifications were applied to the input features. Template models—which AlphaFold 3 uses to infer general structural information and which could artificially inflate mutational invariance—were excluded (Abramson et al., 2024). For both deletion and point mutations, the same random seed was used across all mutation thresholds to ensure comparability of model accuracy. All default settings were used for MSA generation in AlphaFold 3.

In the deletion mutation regime, multiple sequence alignments (MSAs) were modified by removing columns corresponding to deleted residues (see Algorithm 2 for pseudocode). This ensured compatibility with AlphaFold 3, which cannot process MSAs containing sequences whose lengths differ from that of the input sequence. In contrast, for the point mutation regime, MSAs were left unchanged and only the input sequences were mutated, leveraging the fact that MSAs naturally encode organismal sequence variation (see Algorithm 1 for pseudocode).

For deletion mutations, the paired MSA was excluded from the input data. Previous studies have shown that paired MSAs do not improve structural prediction accuracy; thus, they were omitted to simplify the required modifications.

Algorithm 1 MSA Processing for Point Mutations

function PROCESSMUTATIONMSA($unpairedMSA$, $mutation\_map$)
  $first\_sequence \leftarrow unpairedMSA[0]$ {Get query sequence}
  $pos \leftarrow 0$
  $new\_seq \leftarrow$ ""
  for each $char$ in $first\_sequence$ **do**
    if $char$ is uppercase or gap **then**
      if $pos \in mutation\_map$ **then**
        $new\_seq \leftarrow new\_seq + mutation\_map[pos]$
      else
        $new\_seq \leftarrow new\_seq + char$
      end if
      $pos \leftarrow pos + 1$
    else
      $new\_seq \leftarrow new\_seq + char$
    end if
  end for
  $unpairedMSA[0] \leftarrow new\_seq$ {Update first sequence}
  $pairedMSA[0] \leftarrow new\_seq$ {Update first sequence}
  $templates \leftarrow []$
  return $unpairedMSA$, $pairedMSA$, $templates$
end function=0

To verify that modifying MSAs—rather than regenerating them after mutation—did not introduce artificial mutational invariance, MSAs were fully regenerated for the same 200 proteins in the dataset using AlphaFold 3's standard MSA generation pipeline. These proteins were first mutated at all tested thresholds (5%, 10%, 20%, 40%, and 70% for point mutations; 1%, 3%, 5%, and 10% for deletions), and then new MSAs were generated de novo for each mutated sequence using the default AlphaFold 3 MSA generation pipeline. The paired MSA was retained in this regeneration experiment, as it was accurately generated for the mutated sequences. No structural templates were used. No major differences in TM-score or AlphaFold 3 ranking score between models using modified versus regenerated MSAs were noted. Interestingly, models with regenerated MSAs exhibited slightly greater mutational invariance than their directly modified counterparts, suggesting that MSA regeneration only exacerbates the model's structural invariance.

For the comparative analysis between AlphaFold 3 and AlphaFold 2, both models were configured with ten recycles and five samples (Abramson et al., 2024; Jumper et al., 2021). AlphaFold 3 was run without templates to ensure that its template-based features, which are absent in AlphaFold 2, did not confer an unfair performance advantage. As noted in the Results section, this analysis was conducted on 100 proteins drawn from novel bins—50 monomers and 50 multimers—which had no homologs in the AlphaFold 3 training set.

Algorithm 2 MSA Processing for Deletion Mutations

```
function ProcessDeletionMSA(unpairedMSA, deleted_positions)
    for each sequence in unpairedMSA do
        pos ← 0
        new_seq ← ""
        for each char in sequence do
            if char is uppercase or gap then
                if pos ∉ deleted_positions then
                    new_seq ← new_seq + char
                end if
                pos ← pos + 1
            else
                new_seq ← new_seq + char
            end if
        end for
        sequence ← new_seq
    end for
    pairedMSA ← build from final sequence
    templates ← []
    return unpairedMSA, pairedMSA, templates
end function=0
```

B.5. Protein Mutation Details

The point mutations applied to proteins in the 200-protein dataset were designed to be as deleterious as possible, ensuring that the mutations were unlikely to incidentally enhance structural stability. The mapping of these mutations is summarized in the Table 3 in the Appendices. To further increase the likelihood of major structural disruption, the residue mutation algorithms for both point and deletion mutations incorporated a positional bias that preferentially selected residues near the center of the protein (Yu et al., 2024). This bias was implemented mathematically by assigning each residue i a weight

$$w_i = \epsilon + (1 - \epsilon)\left[1 - \frac{|i - (L-1)/2|}{(L-1)/2}\right], \quad i = 0, 1, \ldots, L-1$$

where L is the sequence length and ϵ defines the weight for residues at the sequence termini, which, for the purposes of the study, was set to 0.1. This scheme ensures that central residues, which are more likely to contribute to the protein's core stability, have a higher probability of mutation.

B.6. ESMFold Configuration and Analysis

To compare AlphaFold 3's mutational sensitivity with that of an alternative deep learning–based structure prediction model, we performed parallel analyses using ESMFold (Lin et al., 2023; 2022). The ESMFold implementation was obtained from the *HuggingFace* (Wolf et al., 2020; Jones et al., 2024) repository and run with all default configuration settings. Because ESMFold was not trained to predict multimeric protein structures, this comparison was restricted to monomeric proteins only (Lin et al., 2023). Specifically, we analyzed the 100 monomeric proteins from the 200-protein dataset, comprising 50 from the monomer-novel bin and 50 from the monomer-similar bin. Point mutations and deletion mutations were applied to these proteins using the identical mutation strategy employed for AlphaFold 3. For each mutated protein, ESMFold predictions were compared to the original unmutated ESMFold prediction.

B.7. FoldSeek Exhaustive Structural Search

Each of the 200 experimentally determined structures was queried against FoldSeek's PDB100 database (van Kempen et al., 2023), which comprises approximately 340,000 representative chains clustered at 100% sequence identity from the full

Protein Data Bank. Searches were conducted in exhaustive mode (`--exhaustive-search 1`) with TM-align-based scoring (`--alignment-type 1`), bypassing FoldSeek's heuristic prefilter to guarantee globally optimal structural matches, yielding approximately 9,800 hits per query. Results were post-filtered to retain only hits deposited on or before September 30, 2021, using a list of approximately 180,000 pre-cutoff PDB identifiers obtained from RCSB PDB (noa). The query-normalized TM-score (qtmscore) was used as the primary structural similarity metric, as normalizing by query length prevents partial domain matches from inflating similarity estimates.

For monomeric proteins, AlphaFold 3's monomeric ranking score, pTM, was used as the confidence metric; for multimeric complexes, the ranking score ($0.8 \times$ ipTM $+ 0.2 \times$ pTM) was used (Abramson et al., 2024), jointly capturing fold accuracy and interface quality. For monomers, the top 100 pre-cutoff hits were extracted per query and ranked by qtmscore. One monomer had no pre-cutoff hits, yielding $n = 99$ for the primary correlation analyses; three further monomers had fewer than 100 pre-cutoff hits, yielding $n = 96$ for the neighborhood density analysis. For multimeric complexes, FoldSeek's monomer search mode decomposed each complex into constituent chains, which were searched independently. Per-chain neighborhood statistics were computed from the top 100 pre-cutoff hits and averaged across chains to obtain complex-level estimates. Two multimers were excluded because at least one chain had fewer than 100 pre-cutoff hits, yielding $n = 98$ used throughout. Minimum-across-chains and chain-length-weighted aggregation schemes were also tested and produced comparable results. Only mean aggregation is reported for simplicity.

B.8. Structure Comparison Metrics

TM-score and DockQ were calculated using the OpenStructure package (Biasini et al., 2013). TM-score quantifies global structural similarity between two protein structures and is normalized to a range of 0 to 1, where values above 0.5 typically indicate the same fold (Zhang & Skolnick, 2005). DockQ assesses the quality of predicted protein-protein interfaces by combining information about interface RMSD, ligand RMSD, and native contact fraction (Basu & Wallner, 2016). For multimeric proteins with more than two chains, the weighted average DockQ score across all pairwise chain interfaces was computed and used as the reported metric, with a standard threshold of 0.23 used for a valid interfacial conformation (Basu & Wallner, 2016; Mirabello & Wallner, 2024). Both metrics were calculated by comparing mutated predictions to the original unmutated AlphaFold 3 prediction for the mutational invariance analyses, or to experimental structures from the Protein Data Bank for the model comparison analyses.

B.9. Fold-Switching Protein Mutation Scheme

For the fifteen experimentally validated fold-switching proteins, mutations were applied with a modified position-weighting scheme that prioritized residues or regions identified in previous studies as inducing conformational transitions. The weighting formula combined the center bias used for the main 200-protein dataset with an additional multiplier for experimentally validated regions:

$$w_i = \left[\epsilon + (1-\epsilon)\left(1 - \frac{|i - (L-1)/2|}{(L-1)/2}\right)\right] \times r_i,$$

where $\epsilon = 0.1$ defines the edge penalty and r_i is a range multiplier set to 1.5 for positions within specified mutation ranges and 1.0 otherwise. Discrete positions identified in the literature as critical for fold-switching were treated as mandatory mutations and selected first before random sampling, ensuring that experimentally validated switch-inducing residues were always mutated while residues in broader functional regions received a 50% increased probability of selection. Specific mutation ranges and discrete positions for each protein are listed in Table 4.

Metric	Correlation	Coefficient	5%	10%	20%	40%	70%	Mean
TM-score	Pearson	−0.67	0.74	0.69	0.54	0.34	0.27	0.52
	Spearman	−0.73			—			

Table 2. Relationship between mutation percentage and structural prediction accuracy for fold-switching proteins under ESMFold. TM-score correlation with mutation level and mean values at each threshold, computed across fifteen experimentally validated fold-switching proteins from Porter et al. (Porter & Looger, 2018).

From	To	From	To	From	To	From	To
S	W	N	A	K	G	F	T
T	F	Q	G	R	A	M	Q
C	W	Y	G	W	S	L	N
D	W	H	A	A	Q	I	N
E	F	G	K	V	E	P	R

Table 3. Point mutation transformation pairs used for maximal structural disruption. Standard single-letter abbreviations denote the 20 canonical amino acids.

PDB (Chain A)	Predicted in (Porter & Looger, 2018)	Manual Search	Citation(s)
2KXO	1–89	2–30, 24, 25	(Park et al., 2011; Ayed et al., 2017)
2LSH	29–115	—	(Morris et al., 2013)
4OV8	247–318	128–147, 298–313	(Lukoyanova et al., 2015)
2MZ7	267–312	275–280, 306–311	(Kadavath et al., 2015)
4PMK	27–62	—	(Hamiaux et al., 2014)
2N4O	16–69	—	(Pham et al., 2016)
2KTM	167–201	182–217	(Adrover et al., 2010)
2LE3	N/A	19–30, 24	(Rao et al., 2011a;b)
2X9C	N/A	69–70, 71–76	(Poyraz et al., 2010)
3J9E	2–71	1–68, 69–354	(Zhang et al., 2016)
5SUZ	474–509, 415–509	436; 442–447, 499, 460	(Gammons et al., 2016)
4HLS	146–222	170, 174	(Sweeting et al., 2013)
1S5P	48–107, 98–189, 208–274	—	(Zhao et al., 2004)
3TKA	236–313	—	(Wei et al., 2012)
3GAX	48–120	43–59	(Szymańska et al., 2012)

Table 4. Residues and regions associated with fold switching for each protein in the fold-switching dataset, as collated by Porter et al. (Porter & Looger, 2018) and identified via manual literature search. Entries marked with dashes (—) indicate that no precise fold-switching region was identified.

	Metric	Pearson r	p-value	Spearman ρ	p-value
Monomers	Top 1	0.320	1.5×10^{-3}	0.352	4.4×10^{-4}
	Top 5	0.324	1.3×10^{-3}	0.333	9.2×10^{-4}
	Top 10	0.296	3.5×10^{-3}	0.308	2.3×10^{-3}
	Top 20	0.253	0.013	0.269	8.0×10^{-3}
	Top 100	0.259	0.011	0.317	1.6×10^{-3}
	Drop (1→100)	0.037	0.722	−0.038	0.715
Multimers	Top 1	0.268	7.7×10^{-3}	0.296	3.1×10^{-3}
	Top 5	0.278	5.7×10^{-3}	0.360	2.7×10^{-4}
	Top 10	0.294	3.3×10^{-3}	0.395	5.8×10^{-5}
	Top 20	0.333	8.1×10^{-4}	0.430	9.9×10^{-6}
	Top 100	0.382	1.1×10^{-4}	0.457	2.3×10^{-6}
	Drop (1→100)	−0.168	0.099	−0.153	0.133

Table 5. Correlation between AlphaFold 3 confidence and structural template neighborhood statistics for monomers ($n = 96$) and multimers ($n = 98$). Top N refers to the mean query-normalized TM-score of the N best pre-cutoff structural matches from FoldSeek. Confidence is measured by AlphaFold 3 ranking score.

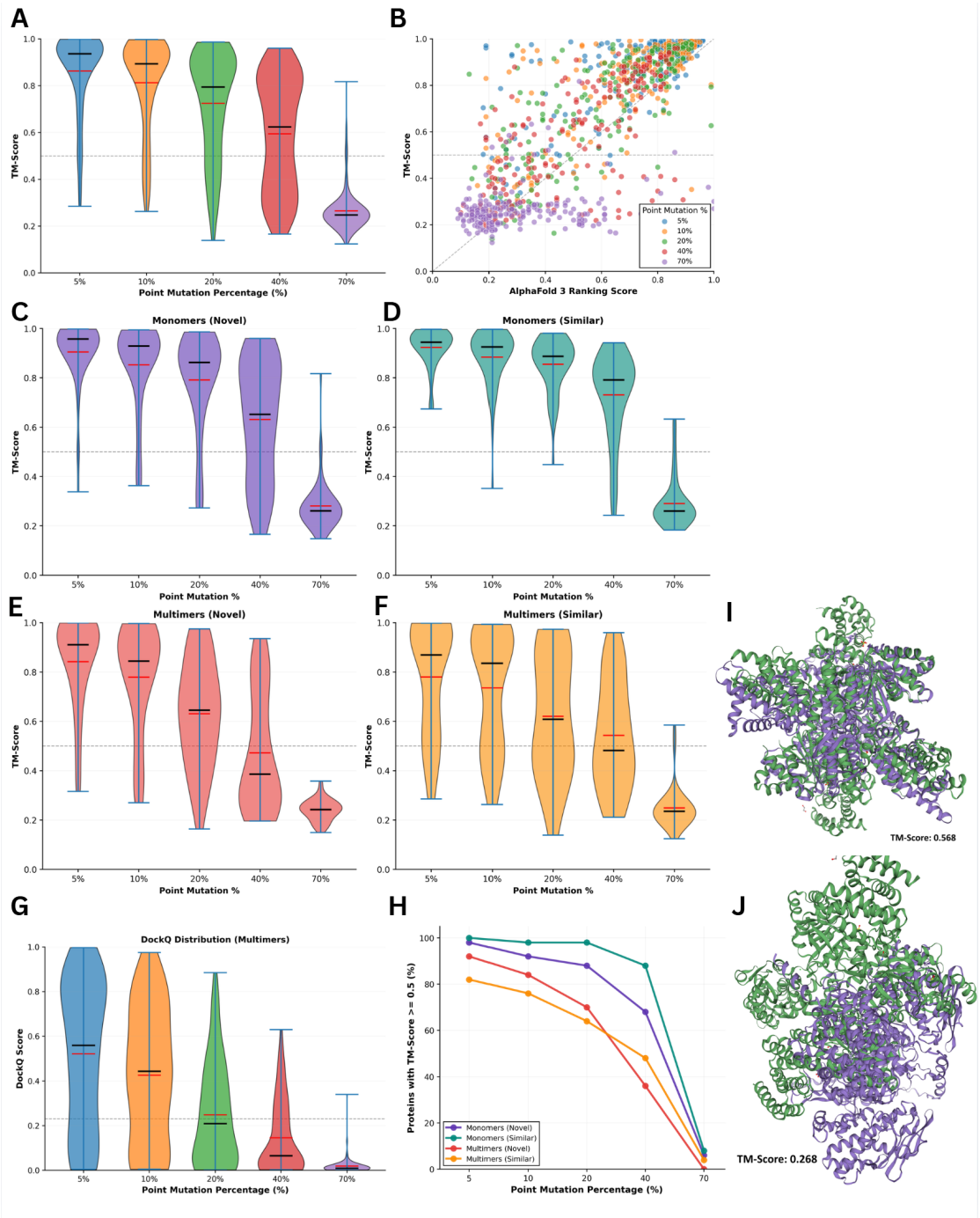

Figure 1. **AlphaFold 3 exhibits structural invariance under adversarial point mutations.** **A** Violin plots showing TM-score distributions for the full 200-protein dataset at each mutation threshold (5%, 10%, 20%, 40%, 70%), comparing mutated predictions to unmutated AlphaFold 3 predictions. **B** Correlation between AlphaFold 3 ranking confidence scores and structural accuracy (TM-score). Each point represents a single protein, colored by mutation percentage. **C–F** TM-score distributions stratified by protein category: (**C**) monomer-novel, (**D**) monomer-similar, (**E**) multimer-novel, and (**F**) multimer-similar. **G** DockQ score distributions for multimeric proteins (novel and similar bins combined), showing interfacial accuracy degradation across mutation thresholds. **H** Survival curve indicating the fraction of proteins maintaining accurate global fold (TM-score ≥ 0.5) at each mutation level. **I** Structural superposition of 7OCN with 40% mutation (purple) onto the unmutated prediction (green), showing preserved fold architecture (TM-score = 0.568). **J** Structural superposition of 7OCN with 70% mutation (purple) onto the unmutated prediction (green), showing fold destruction (TM-score = 0.268). All TM-scores and DockQ values reflect comparison to the unmutated AlphaFold 3 prediction rather than experimental structures.

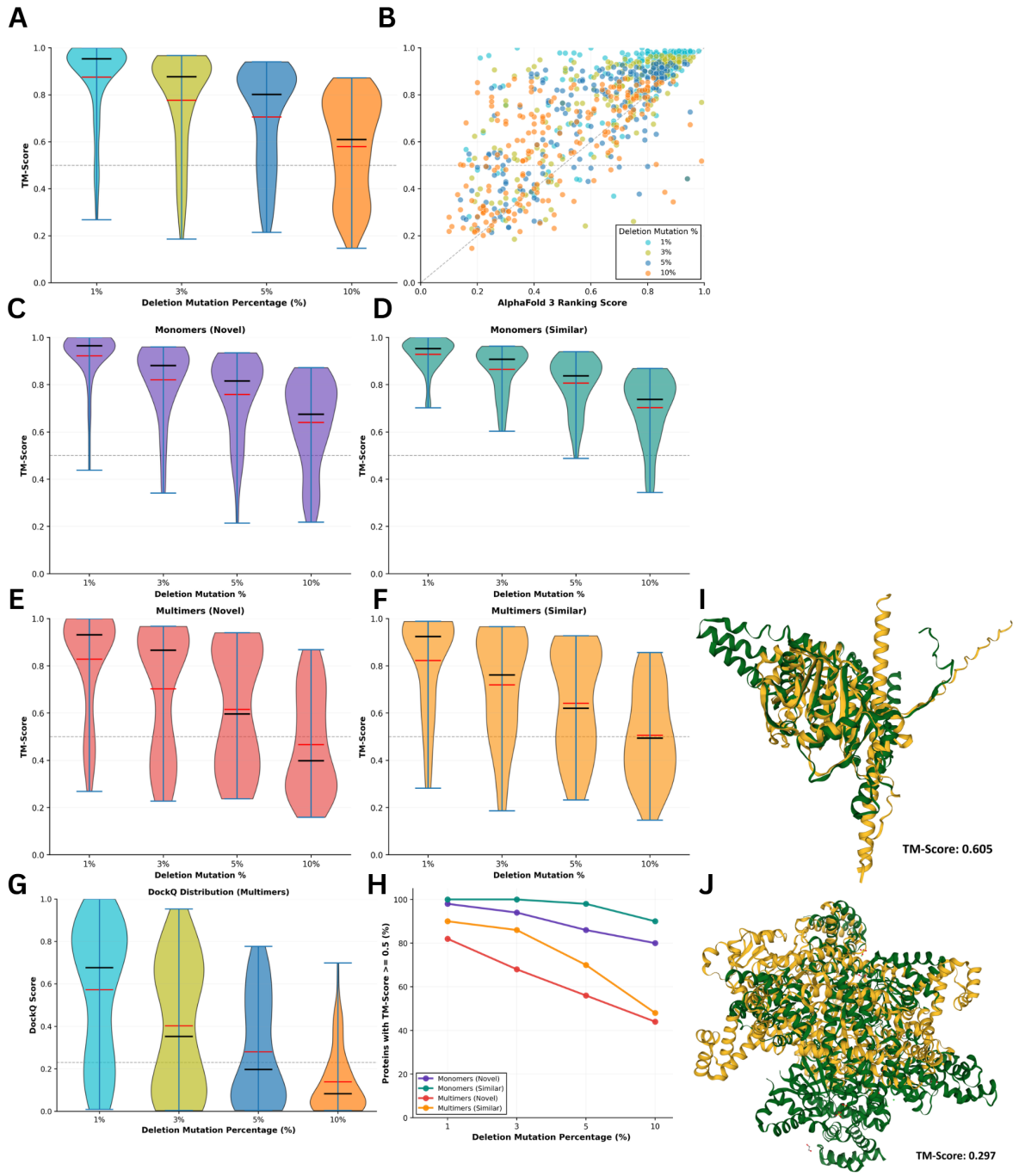

Figure 2. **AlphaFold 3 exhibits structural invariance under adversarial deletion mutations.** **A** Violin plots showing TM-score distributions for the full 200-protein dataset at each deletion threshold (1%, 3%, 5%, 10%), comparing mutated predictions to unmutated AlphaFold 3 predictions. **B** Correlation between AlphaFold 3 ranking confidence scores and structural accuracy (TM-score). Each point represents a single protein, colored by deletion percentage. **C–F** TM-score distributions stratified by protein category: (**C**) monomer-novel, (**D**) monomer-similar, (**E**) multimer-novel, and (**F**) multimer-similar. **G** DockQ score distributions for multimeric proteins (novel and similar bins combined), showing interfacial accuracy degradation across deletion thresholds. **H** Survival curve indicating the fraction of proteins maintaining accurate global fold (TM-score ≥ 0.5) at each deletion level. **I** Structural superposition of 8CQZ at 10% deletion (yellow) onto the unmutated prediction (green), showing preserved fold architecture (TM-score = 0.605). **J** Structural superposition of 7OCN at 10% deletion (yellow) onto the unmutated prediction (green), showing fold destruction (TM-score = 0.297). All TM-scores and DockQ values reflect comparison to the unmutated AlphaFold 3 prediction rather than experimental structures.

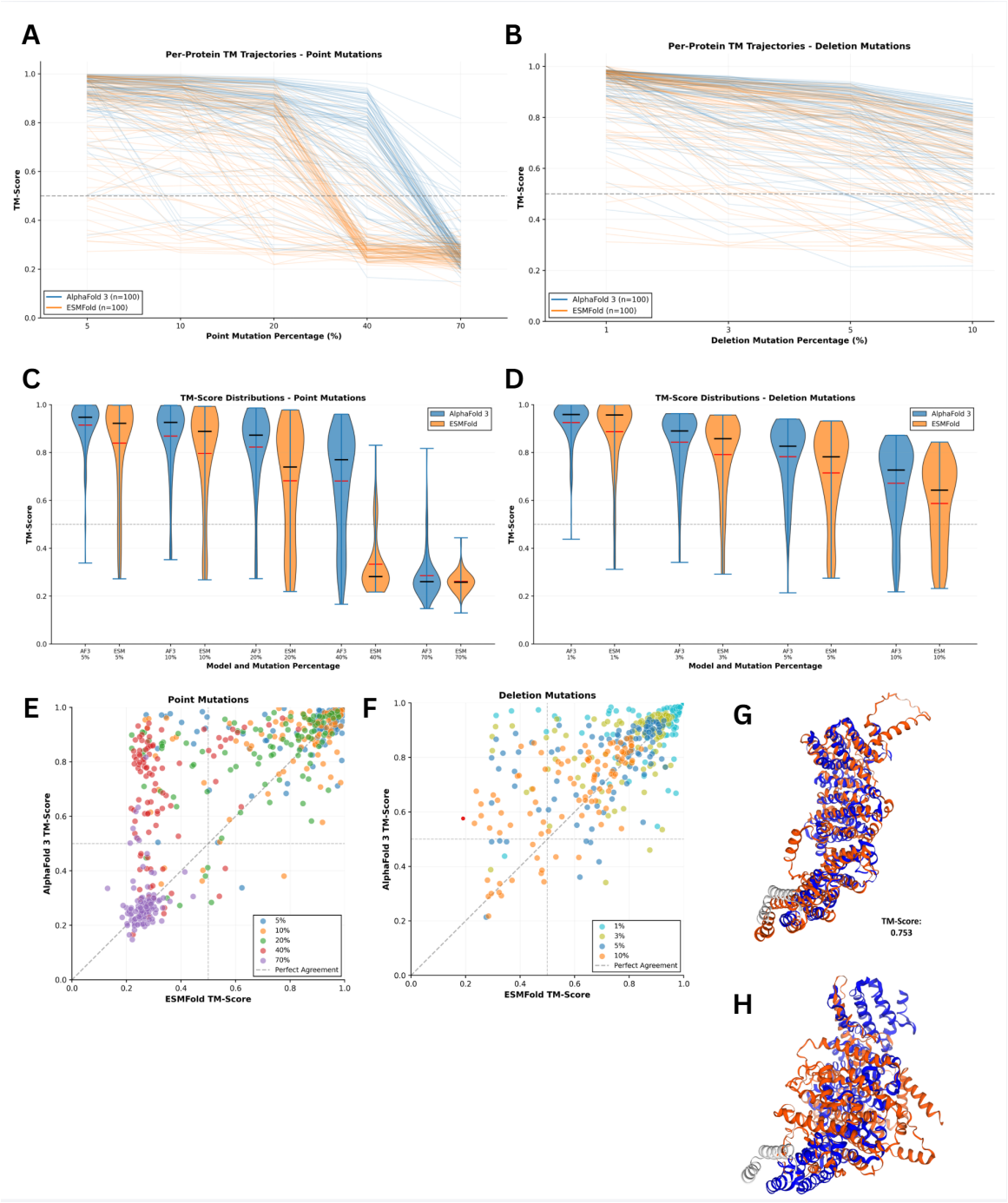

Figure 3. **ESMFold exhibits greater sensitivity to point mutations than AlphaFold 3.** **A** Individual protein accuracy trajectories under point mutations. Each line represents a single protein, with ESMFold (orange) and AlphaFold 3 (blue) predictions compared to their respective unmutated structures. **B** Individual protein accuracy trajectories under deletion mutations, colored as in **A**. **C** TM-score distributions across point mutation thresholds for ESMFold (orange) and AlphaFold 3 (blue). **D** TM-score distributions across deletion mutation thresholds, colored as in **C**. **E** Direct comparison of model accuracy under point mutations. Each point represents a protein at a given mutation threshold, with color indicating mutation percentage. Points above the diagonal indicate greater accuracy for AlphaFold 3. **F** Direct comparison of model accuracy under deletion mutations, displayed as in **E**. **G** Structural superposition of AlphaFold 3 predictions for protein 8RQT: 40% mutated structure (red) aligned to unmutated prediction (blue), demonstrating structural preservation (TM-score = 0.753). **H** Structural superposition of ESMFold predictions for protein 8RQT: 40% mutated structure (red) aligned to unmutated prediction (blue), demonstrating fold destruction (TM-score = 0.293). All comparisons are between mutated and unmutated predictions from the same model.

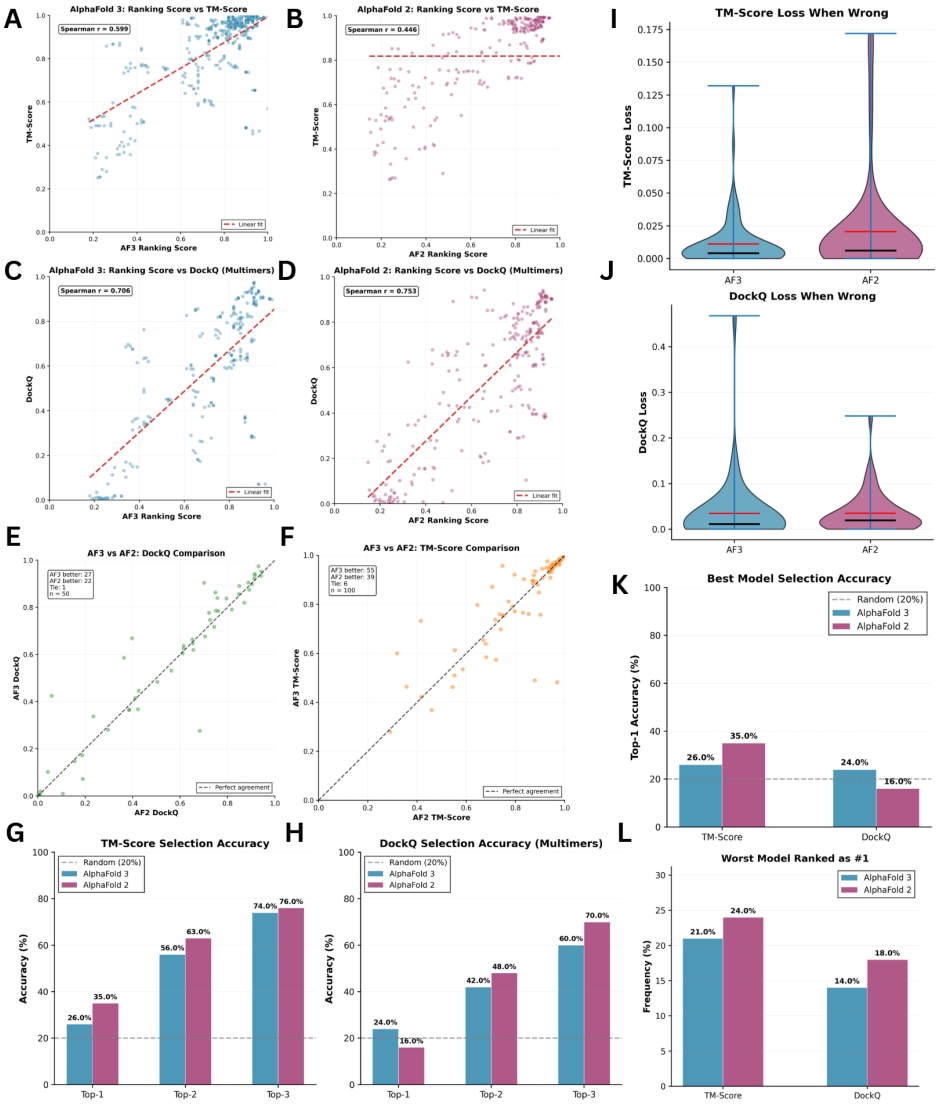

Figure 4. **AlphaFold 2 and AlphaFold 3 confidence metrics fail to reliably identify the most accurate predictions. A–D** Correlation between ranking confidence scores and structural accuracy for AlphaFold 3 (**A**, **C**) and AlphaFold 2 (**B**, **D**). TM-score is shown for monomers (**A**, **B**) and DockQ for multimers (**C**, **D**). Spearman correlation coefficients are indicated, with lines of best fit shown in red. **E–F** Direct comparison of structural accuracy between AlphaFold 2 and AlphaFold 3 on novel proteins, measured by (**E**) TM-score (monomers and multimers) and (**F**) DockQ (multimers only). Points above the diagonal indicate superior AlphaFold 3 performance. **G** Frequency with which each model selects the most accurate structure (rank 1), or a structure within the top 2 or top 3 most accurate, based on TM-score across 100 proteins. AlphaFold 3 (blue) and AlphaFold 2 (purple). **H** Model selection accuracy based on DockQ for 50 multimeric proteins, displayed as in **G**. **I** Distribution of TM-score loss incurred when the model-selected structure is not the most accurate prediction. Larger values indicate more severe selection errors. **J** Distribution of DockQ loss incurred by incorrect model selection, displayed as in **I**. **K** Combined view of model selection accuracy for both TM-score and DockQ metrics, showing the percentage of proteins for which each model correctly identifies the best prediction. **L** Frequency with which each model selects the least accurate (worst) prediction, highlighting severe ranking failures. Both models exhibit poor calibration, selecting the optimal structure in only 16–35% of cases.