Hallucinations in AlphaFold 3 for Intrinsically Disordered Proteins with disorder in Biological Process Residues

Anonymous Author(s)
Affiliation
Address
email

Abstract

Protein structure prediction has advanced significantly with the introduction of AlphaFold3, a diffusion-based model capable of predicting complex biomolecular interactions across proteins, nucleic acids, small molecules, and ions. While AlphaFold3 demonstrates high accuracy in folded proteins, its performance on intrinsically disordered proteins (IDPs)—which comprise 30–40 percent of the human proteome and play critical roles in transcription, signaling, and disease—remains less explored. This study evaluated AlphaFold3's predictions of IDPs with a focus on intrinsically disordered regions (IDRs) using 72 proteins curated from the DisProt database. Predictions were generated across multiple random seeds and ensemble outputs, and residue-level pLDDT scores were compared with experimental disorder annotations. Our analysis reveals that 32 percent of residues are misaligned with DisProt, with 22 percent representing hallucinations where AlphaFold3 incorrectly predicts order in disordered regions or vice versa. Additionally, 10 percent of residues exhibited context-driven misalignment, suggesting that AlphaFold3 implicitly incorporates stable structural assumptions. Importantly, 18 percent of residues associated with biological processes showed hallucinations, raising concerns about downstream implications in drug discovery and disease research. These findings highlight the limitations of AlphaFold3 in modeling IDRs, the need for refined hallucination metrics beyond the pLDDT, and the importance of integrating experimental disorder data to improve the prediction reliability.

1 Introduction

1.1 Significance of Protein Structure Prediction and AlphaFold 3 advancements

Protein structure prediction is of great importance because of its varied downstream applications in drug discovery and disease studies, along with the economic challenges of using experimental techniques to determine protein structure. AlphaFold3, developed by Google DeepMind in collaboration with Isomorphic Labs, is a transformative advancement in biomolecular (protein) prediction that allows progress in drug design and therapeutics [6]. Advancements include a diffusion-based architecture capable of predicting complex biomolecular interactions, including proteins, nucleic acids, small molecules, ions, and modified residues [1]. AlphaFold is consistently evolving with widespread adoption in the domains of structural biology, and the open-sourcing of AlphaFold3 in November 2024 is a significant contribution to research progress. In addition, its predictions have been consistently validated against experimental studies [8].

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

1.2 Intrinsically Disordered Proteins (IDPs) and Regions (IDRs)

The success of AlphaFold 3 and its earlier versions in folded proteins is discussed while expressing its accuracy; however, the accuracy of intrinsically disordered proteins is limited. Intrinsically disordered proteins (IDP) are proteins which lack a stable three-dimensional structure at physiological conditions but are fully functional. The flexibility in the intrinsically disordered regions (IDRs) of the IDP offers the scope to interact with multiple targets [12]. One of the fundamental aspects of understanding is that disorders are not necessarily binary in nature, and may be represented based on context-dependent diverse properties and functions. These properties, when observed in a specific context, exhibit variability in parameters such as pH, localization, binding, and post-translational modifications [2].

The biological relevance of IDPs and Intrinsically Disordered Regions (IDR) is being studied further, considering their prevalence in the eukaryotic genome (60%). Furthermore, IDP's play an indispensable role in biological processes such as transcription, translation, cell cycle, and signaling, and in turn, diseases. Intrinsically disordered proteins (IDPs), comprising 30–40% of the proteome, are critical in diseases including neuro-degenerative disorders and cancer. For instance, 80% of human cancer-associated proteins have long IDRs (e.g., p53 contains 50% IDR in its sequence) [14][10][12]. [5] exhibited that neither AlphaFold2 nor AlphaFold3 did not fully capture structural characteristics of α-synuclein, an intrinsically disordered protein. Earlier studies have identified that Alphafold 2 demonstrates promising results in IDR, specifically with reference to conditionally folding binding regions/residues [11]. Researchers have expressed some possible limitations in terms of chirality determination, conformation, modelling experimental parameters, and predicting structural flexibility through AF3 [9][7][1]. However, there is no extensive research that specifically analyses AlphaFold3's performance against that of IDR. AlphaFold3 has introduced a diffusion-based probabilistic model, which, while improving the prediction efficiency of many biomolecules and their interactions, has also increased the rate of hallucinations [1][7].

1.3 Research Objectives and Scope

This research attempts to conduct such an analysis of IDP's with a focus on IDR's that have an impact on biological processes, with a focus on hallucinations that arise in predicting disordered regions. Further, given the learnings from AlphaFold2 research studies and AlphaFold3 diffusion-based ensemble architecture, this research attempts to cover implications of reproducibility and accuracy with reference to context-dependent IDRs.

2 Methodology

2.1 DisProt Database

DisProt is a database that links the structural and functional information of intrinsically disordered proteins [2]. DisProt is the primary database manually curated and annotated with experimental results for IDPs and IDRs and has evolved with over 3000 IDPs. DisProt contains annotations relating to functional and structural aspects of disorders like disorder state, structural transition, biological processes, and molecular function based on experimental validation

2.2 Protein Selection Criteria

This research leverages the DisProt database to identify relevant IDPs for the research and associated information. The study selected 72 proteins for this study from the DisProt database by identifying proteins with Bio-Process and having IDRs greater than 75% overall.

2.3 AlphaFold3 Prediction Setup

The AlphaFold server generated predictions for the identified IDPs. In order to understand the five ensemble variabilities of AlphaFold3, we experimented with seed variations to observe the extent of variability in predictions for IDRs. To that end, three types of seeds were provided as input while generating the predictions on the AlphaFold Server as follows: (1) No seed was provided manually. The AlphaFold server automatically attributes a random seed. (2) 5 as seeds and (3)

1234567890 as seeds. This approach resulted in three seed variations and five ensemble outputs for each, thereby having 15 model prediction outputs to evaluate reproducibility across prediction runs for each identified IDP.

2.4 Analysis of pLDDT Scores and Variability

The CIF files from the predictions were used to extract the pLDDT scores from the B-factor field programmatically. The pLDDT is a confidence score that indicates the reliability of the structural predictions for each residue, with scores above 70 typically considered to indicate high confidence [15]. Furthermore, pLDDT scores represent protein disorder with a strong correlation with IDRs [3].

The pLDDT scores were parsed to obtain the residue-level scores for each protein. These scores were compared for each seed to determine the seed-based variation for each protein.

pLDDT scores greater than 70 were considered as 'ordered' and less than 70 were considered as 'disordered' in correlation with the DisProt database. An analysis of the variance in predictions based on seed-based input variations for the identified IDP in determining the variability influenced by different seed-based protein predictions.

2.5 Classification of Predictions

A systemic comparison between the IDRs (at a residue level) for each protein from the DisProt database and the predictions (ordered/disordered using pLDDT scores) supported the determination of how well the AlphaFold3 protein predictions align with the ordered/disordered references from DisProt.

Based on the above analysis, proteins were classified into three types: aligned (when AF3 and DisProt are aligned), hallucination (when AF3 shows order but DisProt shows disorder for regions that are not involved in structural transition, and when AF3 shows disorder while DisProt shows order), and possible context-driven misalignment (when AF3 shows order for experimentally proven disordered residues with structural transition). 45 proteins were not considered for context based due to the absence of the structural transition annotation for these proteins.

2.6 Context-Based Interaction Modeling

Of the 27 proteins with context-driven misalignment, 17 proteins were modelled with other biomolecules with which they typically interact to determine the basis for these hallucinations. The other 10 proteins could not be modelled because the interacting components could not be modelled using AlphaFold3 (e.g., environment-induced disorder).

For the 17 proteins, potential interacting biomolecules were identified (based on structural transition literature available in DisProt) and predictions were gathered (adopting the same 3 seed approach) from AlphaFold server, the associated pLDDT were extracted and were classified as 'ordered' or 'disordered' based on a threshold pLDDT score of 70 (as referred above). A comparative analysis was performed for the prediction results similar to the above process.

3 Results

3.1 Variance based on seed input

The analysis revealed a lack of significant variance in AlphaFold3 predictions when varying seed inputs and using ensemble models, with consistent pLDDT scores across seeds (e.g., no seed, 5, 1234567890). This suggests the ensemble approach may not effectively capture structural variability in IDRs across multiple runs.

3.2 Alignment with DisProt Annotations

Initial analysis of alignment of AlphaFold3 predictions as ordered/disordered was done by comparing the ordered/disordered as documented in DisProt for the identified 72 proteins. More than 50% of the proteins had less than 70% alignment with DisProt. Furthermore, the results showed that 68% of the residues aligned with DisProt and 32% did not.

In addition, 7.6% of the aligned residues were involved in structural transition, which implies that AF3 predicted the native structural state of the residues without assuming an unrepresented context. For instance, Apolipoprotein AI (DP04271) exhibited order in only one residue, against 242 residues (91%) that were recorded as disordered in DisProt.

3.3 Hallucination Analysis

This study identified two types of hallucinations. (a) DisProt shows order in the residue, but AF3 predicts the residue with low confidence, implying disorder. (b) DisProt shows disorder, and AF3 predicts it with high confidence without a structural transition potential.

[1] noted that hallucinations (order) in disordered regions were typically flagged with low confidence scores. However, our study found instances where high confidence scores were assigned to disordered regions and ordered structures were predicted with low confidence, both of which were identified as hallucinations in this study.

22% of the residues were hallucinations ((a)+(b)). For instance, proteins such as the nuclear export protein (DP00871), calreticulin (DP00333), functional amyloid subunits FapB and FapC (DP04435 and DP04433, respectively), and von Hippel-Lindau disease tumor suppressor (DP00287) exhibited hallucinations with more than 70% of residues.

3.4 Context-Driven Misalignment

Instances where AF3 predicted a residue with high confidence when the DisProt data showed disorder and is known to have structural transition were identified for subsequent validations. Context-based interactions were identified, and predictions were run for these proteins on AF3, considering the potential context-assumed alignment against stable experimental evidence. For instance, Disprot id DP04016 exhibited ordered in 82%, while it is an IDP with 100% of residues to be disordered, indicating potential assumption of context. AlphaFold 3 has mentioned that there are potential instances wherein the predictions showed inclinations to represent closed state confirmation even when the native confirmation is in an open state [1].

A total of 27 proteins (10% of the overall residue counts considered for review) were found to have residues with context-assumed hallucinations, of which 17 proteins were modelled with other biomolecules with which they typically interact to determine the basis for these misalignments. For instance, proteins such as seed maturation protein (DP01442) and uncharacterized protein (DP03738) fold due to environment-induced stress conditions, and proteins such as LEA proteins (DP04016, DP01858, DP04018, DP04019) and ICP47 protein (DP04190) interact with small molecules such as ethylene glycol and sodium dodecyl sulfate, which cannot be represented in AF3. Other proteins like temporin - 1Tl (DP03818), interact with lipids that cannot be modelled in AF3. The Tat protein (DP01087) was modelled with an Fab molecule generated from an antigen-induced mouse, for which the antibody sequence could not be retrieved.

Of the remaining 17 proteins, a few proteins paratox (DP04243) and temporin - 1Tl (DP03818) showed 100% possibility of context-assumed hallucinations, followed by proteins such as antitoxin phd (DP00288), seed maturation protein (DP01442), and Late Embryogenesis Abundant (LEA) family protein (DP04019), indicating possible order due to inherently context-assumed predictions (>80%). [1] suggests that the AF3 process takes into account possible stable confirmation at residue level and attempts to align the folds to form such stabler confirmations.

Comparison of predictions (ordered/disordered) in the native state with predictions in the context-based state from AF3 revealed that the latter results aligned with the former in 89% of residues (15 out of 17 proteins with more than 80% alignment with the previous run). The only exceptions were alpha synuclein (DP00070) with a match of 38% and tat protein (DP00929) with an alignment of 73% of residues. This raises questions on whether context-based consideration has any significant impact on proteins. Alternatively, it also brings into question whether AF3 predictions are inherently influenced by context assumptions.

Comparison of predictions (ordered/disordered) based on DisProt with predictions with context-based state from AF3 revealed that the later results aligned with the former in 68% of residues. Six out of 17 proteins considered for this analysis had over 80% of residues aligned with DisProt, and in

11 out of 17 cases, 65% of residues aligned with DisProt, questioning if the prior predictions were influenced by context-based considerations that AF3 takes into account.

The exceptions to this trend were the following proteins: antitermination protein N (DP00005), minor curli subunit (DP03852), major curli subunit (DP03853), cholera enterotoxin subunit A (DP00250), and paratox (DP04243). While this review and analysis represents the study of whole protein-related predictions, a specific study focused on IDRs may provide more insights into specific proteins across their context-based interactions. However, the research also noted a protein alpha synuclein (DP00070) that exhibited increased alignment with DisProt in the Context-based prediction run, indicating that the context-based interaction made the protein prediction even more disordered than the prior native AF3 run. This contradicts the role of interacting proteins in transitioning disordered residues to order [4]. Another instance of protein interaction is that of the cholera enterotoxin subunit A and its interaction with protein disulfide isomerase (PDI), which functions as a dual agent to disassemble and reassemble the enterotoxin subunits based on the redox environment [13]. Such instances cannot be modelled in the AlphaFold server currently. Our analysis showed an alignment of 84% with the native state AF3 run and 21% with DisProt, which might suggest an inclination toward the oxidation state of PDI, which leads to folding of the enterotoxin by AF3.

3.5 Impact on Biological Processes

This study utilized the biological process annotation available in the DisProt database to determine the impact of hallucinations in the residues experimentally validated to contribute to biological processes. Overall, 18% of the residues involved in the biological processes showed hallucinations.

Of the 72 proteins, 30% did not show hallucinations at segments known to have a biological context. Eighteen % of the proteins showed hallucinations, ranging from 1 to 10%. 8% of the proteins show a higher degree of hallucination (60-80%) at the regions involved in biological processes. For instance, proteins that are functional amyloid subunits, FapB and FapC (DP04435 and DP04433), show 75% and 77% hallucinations, respectively. The von Hippel-Lindau disease tumor suppressor shows 73% hallucinations in residues attributed to biological processes. Protein Vpr (DP03543), an active regulator of SIRT1 (DP03988), and GRASP65 homolog protein 1 (DP02544) were found to have hallucinations in the range of 60% to 65% at biologically relevant residues.

4 Discussion

4.1 Implications of Hallucinations and Misalignments

The human proteome has approx 20000 protein-coding genes (based on current UniProt/Ensembl data) in its canonical form. Of which 30-40% of the proteomes (approximately 8000) are intrinsically disordered. The number of IDP's documented in DisProt is approximately 3000; hence, the above analysis needs to be considered from the lens of both known IDPs with research and annotations on experimental validations and other IDPs without experimental confirmations (remaining approximately 5000).

Hallucinations and misalignments in known IDP's are comparable and validatable in downstream uses including drug discovery, disease diagnostics, and molecule research. However, such opportunities to identify or analyze IDPs without experimental confirmation may not clearly exist. Hence, the learnings and observations from known IDPs provide referrable guidance for downstream considerations of IDPs without experimental confirmation. From this lens, hallucination and misalignment, in general, could lead to misleading diagnostics or drug targets, delaying therapies and inflating costs due to the need for more experiments.

Specifically, hallucinations in known IDPs represent the need for specific metrics to measure hallucinations or identify potential hallucinations. While pLDDT is considered representative for this purpose, a high pLDDT in disordered regions [1] expresses overconfidence of the prediction model, exploring alternative metrics to measure hallucination effectively in such predictions, as we evolve the process of documenting experimental validation of IDPs. In addition, such hallucinations and misalignment may misguide therapeutic targets and biomarker discovery approaches as part of drug discovery or disease-target explorations and diagnostics. In addition, hallucinations and misalignments, especially in IDPs where experimental confirmations are fewer or need more research, leave severe implications in downstream adoption for target identification for drugs.

4.2 Limitations of Current Analysis

The generalizability of the study is constrained by its limited sample size (72 out of 3000+ DisProt proteins) and exclusive focus on DisProt-annotated proteins, omitting uncharacterized intrinsically disordered proteins (IDPs). Additionally, the reliance on seed-based variability analysis may not fully capture AlphaFold3's ensemble diversity, and AlphaFold3's inability to model certain interactions (e.g., lipids and environment-induced effects) poses a significant limitation. The simplification of disorder-order classification through pLDDT thresholds (>70) requires revision, as generic thresholds for determining ordered or disordered could be misleading contextually.

5 Conclusion

This study quantitatively assessed AlphaFold3's performance on intrinsically disordered proteins (IDPs), revealing significant variability in predictions, particularly concerning non-deterministic hallucinations and context-driven misalignments. These findings indicate that while AlphaFold3 excels in folded protein prediction, its application to IDPs necessitates refined evaluation metrics beyond pLDDT, given instances of high confidence scores assigned to disordered regions. The observed practical implications of these discrepancies underscore the critical need for enhanced research and improvement of AF3 model feedback or continuous learning processes, especially as we extend these models to novel IDPs without existing experimental validation.

5.1 Future Directions for IDP Prediction

Future research should aim to expand the dataset to include a wider range of IDPs and intrinsically disordered regions (IDRs) for robust validation. Developing metrics beyond pLDDT is crucial for accurately detecting and quantifying hallucinations in IDP predictions. Integrating multi-omics data, such as pH and post-translational modifications, will enable context-aware predictions, and the enhancement of AlphaFold3 with iterative feedback from DisProt is an essential step towards reducing predictive hallucinations.

5.2 Recommendations for Model Improvement

Given that the ensemble approach did not provide significant variance across multiple seed runs for the identified IDPs, it may be necessary to understand how the ensemble approach could be used to identify hallucinations rather than collectively confident ensembles leading to hallucinations. In addition, a mechanism to allow for AF3 to perform such comparisons leveraging DisProt to enable the opportunity to improve model performance over period.

References

[1] J. Abramson et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, jun 2024.

[2] M. C. Aspromonte et al. Disprot in 2024: improving function annotation of intrinsically disordered proteins. *Nucleic Acids Res*, 52(D1):D434–D441, jan 2024.

[3] A. Bruley, J.-P. Mornon, E. Duprat, and I. Callebaut. Digging into the 3d structure predictions of alphafold2 with low confidence: Disorder and beyond. *Biomolecules*, 12(10):1467, oct 2022.

[4] D. Cartelli, A. Aliverti, A. Barbiroli, et al. α-synuclein is a novel microtubule dynamase. *Scientific Reports*, 6:33289, 2016.

[5] O. Coskuner-Weber. Structures prediction and replica exchange molecular dynamics simulations of α-synuclein: A case study for intrinsically disordered proteins. *Int J Biol Macromol*, 276:133813, sep 2024.

[6] D. Desai, S. V. Kantliwala, J. Vybhavi, R. Ravi, H. Patel, and J. Patel. Review of alphafold 3: Transformative advances in drug design and therapeutics. *Cureus*, jul 2024.

[7] Z. Fang et al. Alphafold 3: an unprecedent opportunity for fundamental research and drug development. *Precis Clin Med*, 8(3), jun 2025.

[8] O. Kovalevskiy, J. Mateos-Garcia, and K. Tunyasuvunakool. Alphafold two years on: Validation and impact. *Proceedings of the National Academy of Sciences*, 121(34), aug 2024.

[9] M. G. Krokidis et al. Alphafold3: An overview of applications and performance insights. *Int J Mol Sci*, 26(8):3671, apr 2025.

[10] F. Lermyte. Roles, characteristics, and analysis of intrinsically disordered proteins: A minireview. *Life*, 10(12):320, nov 2020.

[11] D. Piovesan, A. M. Monzon, and S. C. E. Tosatto. Intrinsic protein disorder and conditional folding in alphafolddb. *Protein Science*, 31(11), nov 2022.

[12] M. Sato. Biological significance of intrinsically disordered protein structure. *Chem-Bio Informatics Journal*, 22(0):26–37, may 2022.

[13] B. Tsai, C. Rodighiero, W. I. Lencer, and T. A. Rapoport. Protein disulfide isomerase acts as a redox-dependent chaperone to unfold cholera toxin. *Cell*, 104(6):937–948, 2001.

[14] S. Wallin. Intrinsically disordered proteins: structural and functional dynamics. *Res Rep Biol*, 8:7–16, feb 2017.

[15] C. J. Williams, V. B. Chen, D. C. Richardson, and J. S. Richardson. Categorizing prediction modes within low-plddt regions of alphafold2 structures. *bioRxiv*, jun 2025.

[1] D. Desai, S. V Kantliwala, J. Vybhavi, R. Ravi, H. Patel, and J. Patel, "Review of AlphaFold 3: Transformative Advances in Drug Design and Therapeutics," Cureus, Jul. 2024, doi: 10.7759/cureus.63646.

[2] J. Abramson et al., "Accurate structure prediction of biomolecular interactions with AlphaFold 3," Nature, vol. 630, no. 8016, pp. 493–500, Jun. 2024, doi: 10.1038/s41586-024-07487-w.

[3] A. Elofsson, "AlphaFold3 at CASP16," Proteins: Structure, Function, and Bioinformatics, Aug. 2025, doi: 10.1002/prot.70044.

[4] O. Kovalevskiy, J. Mateos-Garcia, and K. Tunyasuvunakool, "AlphaFold two years on: Validation and impact," Proceedings of the National Academy of Sciences, vol. 121, no. 34, Aug. 2024, doi: 10.1073/pnas.2315002121.

[5] M. Sato, "Biological Significance of Intrinsically Disordered Protein Structure," Chem-Bio Informatics Journal, vol. 22, no. 0, pp. 26–37, May 2022, doi: 10.1273/cbij.22.26.

[6] M. C. Aspromonte et al., "DisProt in 2024: improving function annotation of intrinsically disordered proteins," Nucleic Acids Res, vol. 52, no. D1, pp. D434–D441, Jan. 2024, doi: 10.1093/nar/gkad928.

[7] S. Wallin, "Intrinsically disordered proteins: structural and functional dynamics," Res Rep Biol, vol. Volume 8, pp. 7–16, Feb. 2017, doi: 10.2147/RRB.S57282.

[8] F. Lermyte, "Roles, Characteristics, and Analysis of Intrinsically Disordered Proteins: A Minireview," Life, vol. 10, no. 12, p. 320, Nov. 2020, doi: 10.3390/life10120320.

[9] O. Coskuner-Weber, "Structures prediction and replica exchange molecular dynamics simulations of α-synuclein: A case study for intrinsically disordered proteins," Int J Biol Macromol, vol. 276, p. 133813, Sep. 2024, doi: 10.1016/j.ijbiomac.2024.133813.

[10] D. Piovesan, A. M. Monzon, and S. C. E. Tosatto, "Intrinsic protein disorder and conditional folding in AlphaFoldDB," Protein Science, vol. 31, no. 11, Nov. 2022, doi: 10.1002/pro.4466.

[11] M. G. Krokidis et al., "AlphaFold3: An Overview of Applications and Performance Insights," Int J Mol Sci, vol. 26, no. 8, p. 3671, Apr. 2025, doi: 10.3390/ijms26083671.

[12] Z. Fang et al., "AlphaFold 3: an unprecedent opportunity for fundamental research and drug development," Precis Clin Med, vol. 8, no. 3, Jun. 2025, doi: 10.1093/pcmedi/pbaf015.

[13] C. J. Williams, V. B. Chen, D. C. Richardson, and J. S. Richardson, "Categorizing prediction modes within low-pLDDT regions of AlphaFold2 structures," Jun. 07, 2025. doi: 10.1101/2025.06.06.658382.

[14] A. Bruley, J.-P. Mornon, E. Duprat, and I. Callebaut, "Digging into the 3D Structure Predictions of AlphaFold2 with Low Confidence: Disorder and Beyond," Biomolecules, vol. 12, no. 10, p. 1467, Oct. 2022, doi: 10.3390/biom12101467.

[15] K. M. Ruff and R. V. Pappu, "AlphaFold and Implications for Intrinsically Disordered Proteins," J Mol Biol, vol. 433, no. 20, p. 167208, Oct. 2021, doi: 10.1016/j.jmb.2021.167208.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] **Justification**: The abstract and introduction outline the evaluation of AlphaFold3's performance on IDPs, focusing on hallucinations and misalignments, which align with the methods, results, and discussion in the paper (Sections 1, 3, 4).

Question: Have you provided proper citation or derivation for claims or methods that rely on prior work? [Yes] **Justification**: Citations are included for AlphaFold3, DisProt, and related IDP studies (e.g., Abramson et al., 2024; Sato, 2022).

2. Limitations

Question: Does your paper discuss the limitations of the work? [Yes] **Justification**: Limitations, such as small sample size and inability to model certain interactions, are discussed in Section 4.2.

Question: If there are limitations for which mitigations exist, did you discuss how an improved future study could be conducted? [Yes] **Justification**: Future directions, including expanding datasets and developing new metrics, are proposed in Section 4.3.

3. Reproducibility

Question: Is enough information provided to reproduce the experiments (e.g., model details, training procedure, hyperparameters, random seeds)? [Yes] **Justification**: Methods detail the use of DisProt, 72-protein selection, AlphaFold3 server, seed variations, and pLDDT analysis (Section 2). Specific seeds (e.g., 5, 1234567890) were provided.

Question: Are the datasets, code, or instructions to access them provided or described in enough detail to be accessed or recreated? [Yes] **Justification**: The DisProt database is referenced, and protein selection criteria are specified (Section 2.1). AlphaFold3 server usage is described, although code is not provided.

Question: Are the evaluation metrics clearly defined and appropriate for the task? [Yes] **Justification**: pLDDT scores with a threshold of 70 for order/disorder classification are defined and correlated with DisProt annotations (Section 2.4).

Question: Are results reported with appropriate measures of statistical significance, error bars, or other descriptions of uncertainty (when applicable)? [No] **Justification**: Results report percentages (e.g., 32

4. Ethics

Question: Does the research adhere to ethical standards (e.g., regarding human subjects, animals, or potential misuse)? [NA] **Justification**: The study uses computational protein prediction and publicly available DisProt data, involving no human subjects, animals, or direct ethical concerns.

Question: If the work has potential negative societal impacts, are these discussed? [Yes] **Justification**: Potential misguidance in drug discovery and diagnostics due to hallucinations is discussed in Section 4.1.

5. Datasets

Question: Are the datasets used clearly described, and is their licensing discussed (if applicable)? [Yes] **Justification**: DisProt is described as the primary dataset, manually curated for IDPs (Section 2.1). Licensing is not explicitly discussed but implied as publicly accessible.

Question: If a new dataset is introduced, is it described in sufficient detail for others to understand its creation and intended use? [NA] **Justification**: No new dataset is introduced; the study relies on the existing DisProt database.

6. Code

Question: If code is used or released, is the license specified? [NA] **Justification**: No code is released; the study uses the AlphaFold3 server for predictions (Section 2.3).

Question: If code is released, is it accompanied by documentation sufficient to allow others to use it? [NA] **Justification**: No code is released, as the study relies on the AlphaFold3 server.

7. Broader Impact

Question: Does the paper discuss the broader impact of the work, including possible societal implications (positive and negative)? [Yes] **Justification**: The discussion highlights implications for drug discovery and diagnostics, noting risks of misguidance due to hallucinations (Section 4.1).

8. Safely Releasing Models

Question: If a model is released, are steps taken to prevent or mitigate potential misuse? [NA] **Justification**: No model is released; the study evaluates AlphaFold3's predictions.

Question: If a model is released, is its license specified? [NA] **Justification**: No model is released; the study uses the existing AlphaFold3 server.

9. Experiments

Question: Are the computational resources used (e.g., hardware, runtime) described? [No] **Justification**: The paper mentions the AlphaFold3 server but does not specify hardware or runtime details (Section 2.3).

Question: Are experiments conducted with multiple random seeds or other methods to measure variability? [Yes] **Justification**: Three seed variations (none, 5, 1234567890) are used to assess prediction variability (Section 2.3).

Question: Are the results compared to appropriate baselines or prior work? [Yes] **Justification**: Predictions are compared to DisProt annotations as a baseline for IDP accuracy (Section 3.1).