AlphaFold3 Compared to ESMFold and OmegaFold: Defining Performance Boundaries for Protein Structure Prediction

Jade Lascoux[1] and Anuar Badrul[2]

[1]Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia
[2]CESI Engineering School, Toulouse, France
Author for correspondence: J. Lascoux, Email: jade.lascoux@viacesi.fr

Abstract

The recent release of AlphaFold3 raised questions about whether architectural modifications reduced its dependence on multiple sequence alignments (MSAs) compared to MSA-free models like ESMFold and OmegaFold. We systematically compared these three tools across 14 proteins spanning monomers, multi-chain assemblies, orphan proteins, and intrinsically disordered proteins (IDPs).
AlphaFold3 achieved best accuracy on monomers with MSA (80% of cases) but exhibited 71% failure rate on multi-chain assemblies (5/7), with failures occurring even with deep MSA for complex hetero-oligomers. AlphaFold3 also showed systematic overconfidence for IDPs. ESMFold demonstrated exceptional robustness with no catastrophic failures, appropriate confidence calibration for disordered proteins, and 57–850× speed advantage over AlphaFold3. OmegaFold showed competitive performance for structured proteins but calibration issues for disorder prediction.
These findings establish practical guidelines: AlphaFold3 for maximum accuracy on structured monomers with MSA; ESMFold for robustness, disorder analysis, and high-throughput applications; experimental validation essential for AlphaFold3 multi-chain predictions.

Keywords: AlphaFold3, ESMFold, OmegaFold, protein structure prediction, MSA dependence, intrinsically disordered proteins, benchmarking

1 Introduction

Proteins are the molecular machines of life, essential to virtually every biological process from immune defense to DNA repair. Yet despite decades of research, we face a striking imbalance: scientists have identified over 250 million protein sequences, but have experimentally determined the 3D structure for only approximately 230,000 of them (less than 0.1%). This gap is not merely academic. Without knowing a protein's structure, we cannot fully understand how it works, predict how mutations cause disease, or design drugs to target it. Traditional experimental methods like X-ray crystallography can take months or years per protein and cost tens of thousands of dollars. This limitation has made computational structure prediction not just useful, but essential.

AlphaFold2's breakthrough performance at CASP14 (December 2020) and subsequent public release in 2021 [1] marked a paradigm shift in protein structure prediction, achieving near-experimental accuracy through an architecture that combined deep multiple sequence alignments (MSAs) with attention-based neural networks. The Evoformer module processed MSA information to capture evolutionary constraints and residue-residue interactions, enabling remarkably accurate predictions for proteins with deep sequence coverage. However, this dependence on high-quality MSAs posed challenges for orphan proteins, rapidly evolving sequences, and de novo designed proteins lacking evolutionary context. AlphaFold3, released in May 2024, introduced substantial architectural changes including replacement of Evoformer with Pairformer. Whether these modifications successfully reduced MSA dependence (a key limitation for expanding AlphaFold's applicability) remains an open question requiring systematic evaluation.

Concurrent with AlphaFold's development, an alternative approach emerged based on protein language models. ESMFold [3] and OmegaFold [4] exemplify this MSA-independent paradigm, using large-scale language models (ESM-2 with 15 billion

parameters and OmegaPLM with 670 million parameters, respectively) to learn evolutionary patterns directly from sequence data. These models promised substantial advantages: dramatically reduced computational time (predictions in seconds rather than minutes or hours) and potentially superior performance on proteins lacking homologous sequences. However, comprehensive comparison of these MSA-free approaches with AlphaFold3 remains limited, particularly for challenging cases such as orphan proteins, multi-chain assemblies, and intrinsically disordered proteins.

Benchmarking studies have compared AlphaFold2, ESMFold, and OmegaFold on structured proteins with available homologs, demonstrating AlphaFold2's superior accuracy and ESMFold's computational efficiency. However, several protein classes present distinct prediction challenges. Intrinsically disordered proteins (IDPs), which comprise approximately 30% of eukaryotic proteomes, lack stable structure and require appropriate confidence calibration. Orphan proteins from poorly characterized organisms have limited evolutionary context. Multi-chain assemblies without deep MSAs test MSA dependence for interface modeling. With AlphaFold3's 2024 release introducing architectural changes, systematic comparison of AlphaFold3, ESMFold, and OmegaFold across these challenging protein classes is needed to define practical performance boundaries for each approach.

Here, we present a systematic comparison of AlphaFold3, ESMFold, and OmegaFold across 14 carefully selected proteins: monomers with varying MSA availability, multi-chain assemblies, a metagenomic orphan protein (MGYP000261684433), and DisProt-validated intrinsically disordered proteins (α-synuclein, fetal tau isoform). We evaluate structural accuracy (RMSD, TM-score), confidence calibration (pLDDT, disorder fractions), and computational efficiency. Our analysis reveals distinct performance boundaries: AlphaFold3 achieves superior accuracy on monomers with MSA but exhibits critical MSA dependence for multi-chain assemblies and systematic overconfidence for disordered proteins; ESMFold demonstrates exceptional robustness with no catastrophic failures and appropriate confidence calibration; OmegaFold shows competitive performance for structured proteins but calibration inconsistencies for disorder prediction. These findings provide practical guidance for model selection based on protein characteristics and project requirements.

2 Methods

2.1 Dataset

We compiled a benchmark dataset of 14 proteins to evaluate prediction performance across diverse structural and evolutionary contexts, spanning three axes: protein size (43–1281 amino acids), structural complexity (monomers to octamers), and MSA availability (deep, shallow, or absent).

The dataset comprises four categories. **Classic monomers** (5 proteins: 1UBQ, 1QYS, 5L33, 6V67 chain A, 6O35 chain A) with experimental structures enable RMSD validation across varying MSA depths. **Multi-chain assemblies** (7 proteins: 1A3N, 6Z4U, 6M0J, 2EKY, 6V67, 6O35, 7R6X) ranging from dimers to octamers test quaternary structure prediction capabilities. 7R6X (SARS-CoV-2 spike RBD in complex with three Fab antibodies, hetero-7-mer) was included to test prediction of complex hetero-oligomeric assemblies. Note that 6O35 and 6V67 appear in both categories, analyzed as isolated chains and complete assemblies respectively. **Proteins without experimental structures** (3 proteins) include one metagenomic orphan (MGYP000261684433, 421 aa) and two DisProt-validated intrinsically disordered proteins (DP00070: α-synuclein, 140 aa; DP03552: fetal tau isoform, 352 aa). **Computational edge case**: 6VXX (SARS-CoV-2 Spike, 1281 aa $\times$ 3 chains) exceeded GPU memory limits for all three tools and was excluded from analysis.

Protein structures were obtained from RCSB PDB (11 proteins), MGnify (1 protein), and DisProt (2 proteins). Complete dataset details are provided in Table 1.

2.2 Structure Prediction Tools

Three tools were evaluated for protein structure prediction: AlphaFold3, ESMFold, and OmegaFold.

AlphaFold3. Predictions were generated using AlphaFold3 pre-installed on the University of Malaya computational server. Protein sequences were provided in JSON format with modelSeeds=[1,2] to generate two predictions per protein, from which the top-ranked model was selected for analysis. AlphaFold3 performs automatic MSA searches and uniquely supports multi-chain complex prediction, simultaneously modeling all chains and their interfaces in a single pass. Predictions were executed on NVIDIA A100 GPUs with computation

Table 1. Dataset composition and characteristics.

Category	Protein	Chains	Size (aa)	MSA Depth
Monomers	1UBQ	1	76	Deep
	1QYS	1	106	Shallow
	5L33	1	62	None
	6V67*	1	43	None
	6O35*	1	102	None
Multi-chain	1A3N	4	141–146	Deep
	6Z4U	2	97×2	Medium
	6M0J	2	229, 603	Medium
	2EKY	8	100×8	Shallow
	6V67*	2	43×2	None
	6O35*	4	102×4	None
	7R6X	7	215–230	Deep
No Exp. Struct.	MGYP000261684433	1	421	None
	DP00070	1	140	N/A
	DP03552	1	352	N/A
Excluded	6VXX	1	1281	Deep

*6V67 and 6O35 analyzed as both isolated chains and complete assemblies
*7R6X: SARS-CoV-2 spike RBD with three Fab antibodies (hetero-7-mer)

times ranging from 16 to 49 minutes per protein.

ESMFold. ESMFold was installed in a dedicated conda environment (Python 3.9, PyTorch 2.x) with manual configuration of OpenFold dependencies and deactivated pre-compiled CUDA kernels due to server constraints. ESMFold uses the ESM-2 protein language model (15 billion parameters) for MSA-independent prediction. Unlike AlphaFold3, ESMFold processes single sequences and cannot model multi-chain assemblies; for multi-chain proteins, each chain was predicted independently. Predictions were executed on NVIDIA A100 GPUs with computation times of 2–17 seconds per sequence.

OmegaFold. OmegaFold v1.0 was installed from the official GitHub repository in a separate conda environment with manual PyTorch configuration and custom PYTHONPATH management. OmegaFold employs the OmegaPLM language model (670 million parameters) for MSA-free structure prediction. Like ESMFold, OmegaFold operates on individual chains without multi-chain complex modeling capabilities. Predictions were executed on NVIDIA A100 GPUs with computation times of 5–70 seconds per sequence, making OmegaFold the fastest tool for small proteins.

All predictions were performed during October–December 2025. Output files included structure coordinates (CIF format for AlphaFold3, PDB format for ESMFold/OmegaFold) and confidence metrics (pLDDT, PAE, PTM, iPTM).

2.3 Evaluation Metrics

Prediction quality was assessed using multiple complementary metrics.

Structural accuracy. Root Mean Square Deviation (RMSD) quantifies the average distance between corresponding Cα atoms in predicted and experimental structures after optimal superposition, with lower values indicating higher accuracy. RMSD was calculated using PyMOL's `super` command, which performs iterative alignment with rejection of poorly aligned atoms. AlphaFold3 CIF files were converted to PDB format using BioPython for standardization. Template Modeling score (TM-score) evaluates topological similarity normalized by protein length (0–1 scale), where values above 0.5 indicate the same fold.

Confidence metrics. Predicted Local Distance Difference Test (pLDDT) scores estimate per-residue prediction confidence on a 0–100 scale: >90 indicates very high confidence, 70–90 suggests reliable backbone prediction, 50–70 indicates low confidence, and <50 suggests disordered regions. For

AlphaFold3, pLDDT values were extracted from output JSON files. For ESMFold and OmegaFold, per-residue pLDDT scores stored in PDB B-factor columns were extracted using BioPython from Cα atoms. Mean pLDDT across all residues was used to assess overall prediction confidence. Additional AlphaFold3 metrics included Predicted TM-score (PTM) and interface PTM (iPTM) for multi-chain assemblies, both ranging from 0 to 1, where values >0.8 indicate high confidence.

Disorder prediction. The fraction of residues classified as disordered (pLDDT < 50) was calculated for each prediction and compared against DisProt experimental annotations for DP00070 and DP03552.

Computational efficiency. Wall-clock time from sequence submission to structure output was recorded via timestamps in execution scripts, including I/O and MSA generation for AlphaFold3.

Multi-chain protein handling. For ESMFold and OmegaFold, individual chains were predicted and evaluated separately. For AlphaFold3, RMSD was calculated both on individual chains (for comparison with ESMFold/OmegaFold) and on complete assemblies (to evaluate quaternary structure prediction).

Predicted structures were visualized using ChimeraX with pLDDT-based color coding to identify well-predicted (dark blue, high pLDDT) and uncertain (red/orange, low pLDDT) regions. For proteins without experimental structures (MGYP, DP00070, DP03552), only confidence metrics and disorder fractions were evaluated.

3 Results

3.1 Monomer Performance Across MSA Depths

Performance on five single-chain proteins with experimental structures demonstrated AlphaFold3's superior accuracy in the majority of cases (Table 2). AlphaFold3 achieved best accuracy on 4 of 5 proteins (80%), with RMSD values ranging from 0.272 Å (5L33) to 0.469 Å (1QYS). ESMFold achieved best accuracy on 1 protein (6V67, 0.492 Å). All three tools produced high-quality predictions with RMSD < 1 Å across all monomers.

For proteins with deep or shallow MSA availability (1UBQ, 1QYS), AlphaFold3 demonstrated advantages over MSA-free models: 0.320 Å versus 0.342 Å (ESMFold) and 0.327 Å (OmegaFold) for 1UBQ; 0.469 Å versus 0.581 Å (ESMFold) and 0.602 Å (OmegaFold) for 1QYS. For proteins without MSA (5L33, 6V67 chain A, 6O35 chain A), performance differences narrowed. AlphaFold3 achieved 0.272 Å and 0.424 Å on 5L33 and 6O35 respectively, compared to ESMFold's 0.299 Å and 0.509 Å. ESMFold outperformed AlphaFold3 on 6V67 (0.492 Å versus 0.645 Å). Computational efficiency is analyzed in Section 3.4.

Table 2. Performance on classic monomers.

Protein	Size (aa)	MSA	AF3 (Å)	ESM (Å)	OF (Å)
1UBQ	76	Deep	**0.320**	0.342	0.327
1QYS	106	Shallow	**0.469**	0.581	0.602
5L33	62	None	**0.272**	0.299	0.435
6V67*	43	None	0.645	**0.492**	0.825
6O35*	102	None	**0.424**	0.509	0.522

*Chain A only. Bold = best RMSD.

3.2 Multi-Chain Assembly Prediction

Multi-chain proteins presented a methodological challenge: AlphaFold3 models complete assemblies while ESMFold and OmegaFold, due to architectural limitations, predict individual chains only. Results revealed distinct performance patterns between these approaches (Table 3).

Individual chain predictions. ESMFold and OmegaFold produced high-quality predictions for isolated chains. ESMFold achieved best accuracy on 4 of 7 proteins (1A3N: 0.297 Å, 2EKY: 0.311 Å, 6V67: 0.492 Å, 6O35: 0.509 Å). OmegaFold achieved best accuracy on 2 proteins: 6M0J (1.011 Å) and 7R6X (0.308 Å). AlphaFold3 achieved best chain accuracy on 6Z4U (0.597 Å). ESMFold demonstrated superior robustness across diverse multi-chain contexts with four of seven predictions below 1 Å RMSD.

Assembly prediction failures. AlphaFold3 exhibited catastrophic failures on 5 of 7 multi-chain assemblies (71%): 2EKY (octamer, 26.067 Å), 6V67 (dimer, 8.746 Å), 6O35 (tetramer, 13.645 Å), 6M0J (heterodimer, 27.765 Å), and 7R6X (hetero-7-mer, 34.098 Å). While most failures occurred on proteins with shallow or absent MSA, 7R6X demonstrates that even deep MSA does not guarantee successful assembly prediction for complex hetero-

oligomers. AlphaFold3 succeeded only on simpler homo-oligomeric assemblies: 6Z4U (0.597 Å) and 1A3N (0.480 Å), suggesting that both MSA depth and assembly complexity determine prediction success.

Quaternary structure paradox. When individual chains were extracted from failed AlphaFold3 assembly predictions and evaluated separately, they exhibited accurate monomer structures. AlphaFold3 thus correctly predicted individual chain folding but failed to assemble them properly in 3D space, indicating that failures stemmed from interface modeling rather than monomer structure prediction.

Confidence calibration issues. Failed assembly predictions exhibited inappropriately high confidence scores (pLDDT 73–84, PTM 0.34–0.70). Notably, 7R6X achieved pLDDT 80–82 despite a catastrophic RMSD of 34.098 Å, though its low iPTM (0.25–0.31) partially detected the interface failure. This systematic overconfidence indicates that AlphaFold3 confidence metrics do not reliably detect quaternary structure prediction failures, particularly for complex hetero-oligomers.

Table 3. Performance on multi-chain proteins.

Prot.	Type	MSA	AF3 (cplx)	ESM (chain)	OF (chain)
2EKY	Octamer	Shallow	26.067	**0.311**	0.434
6V67	Dimer	None	8.746	**0.492**	0.825
6O35	Tetramer	None	13.645	**0.509**	0.522
6Z4U	Dimer	Medium	**0.597**	0.809	5.359
6M0J	Heterod.	Medium	27.765	1.679	**1.011**
1A3N	Hemog.	Deep	0.480	**0.297**	0.379
7R6X	Hetero-7	Deep	34.098	0.364	**0.308**

AF3 on complex; ESM/OF on single chain (A or D). Bold = best. AF3 failures: 5/7 (71%).

3.3 Orphan and Disordered Protein Performance

Three proteins lacking experimental structures were evaluated using computational confidence metrics: one metagenomic orphan protein (MGYP000261684433, 421 residues) and two DisProt-validated intrinsically disordered proteins (DP00070: α-synuclein, 140 residues, 100% disordered; DP03552: fetal tau isoform, 352 residues, 100% disordered). Assessment relied on pLDDT scores, disorder fractions, and ranking scores (Table 4).

Orphan protein (MGYP). The three models produced divergent predictions for MGYP, a metagenomic protein without detectable sequence homologs. ESMFold assigned moderate confidence (pLDDT 65.17) with 22% residues predicted as disordered (fraction 0.22), suggesting partially structured regions. In contrast, AlphaFold3 (pLDDT 47.98, disorder fraction 0.76) and OmegaFold (pLDDT 45.33, disorder fraction 0.73) predicted predominantly unstructured protein with approximately 75% disordered residues. Despite divergent disorder predictions, all three models agreed on absence of well-defined stable structure (consensus pLDDT <70), consistent with expectations for orphan proteins from poorly characterized metagenomic sources.

α-synuclein (DP00070). The three models exhibited strikingly different confidence calibration for this canonical intrinsically disordered protein. ESMFold produced appropriately low pLDDT (31.30) with 98% residues classified as disordered (fraction 0.98), accurately reflecting experimental validation. AlphaFold3 exhibited systematic overconfidence with pLDDT 71.83 (a score typical for well-folded proteins) despite correctly detecting 92% disorder fraction. This overconfidence is consistent with recent systematic evaluation showing AlphaFold3 generates structural hallucinations for 22% of experimentally validated disordered residues in DisProt proteins [5]. OmegaFold assigned appropriate low pLDDT (39.60) but severely miscalibrated disorder fraction (2%), indicating fundamental disorder detection limitations.

Fetal tau (DP03552). All three models assigned moderate pLDDT scores (51.60–58.14) for this 352-residue disordered protein, higher than for α-synuclein. AlphaFold3 again exhibited overconfidence (pLDDT 58.14) despite detecting 100% disorder fraction, demonstrating persistence of the hallucination pattern though less severe than for DP00070. ESMFold produced pLDDT 51.60 with 36% disorder fraction. OmegaFold showed pLDDT 55.52 with 37% disorder fraction. The moderately higher pLDDT scores across all models may reflect tau's experimentally observed compaction in solution despite lacking stable structure [6].

Table 4. Performance on orphan and intrinsically disordered proteins.

		AlphaFold3			ESMFold		OmegaFold	
Protein	**Category**	**pLDDT**	**Frac. Dis.**	**Rank**	**pLDDT**	**Frac. Dis.**	**pLDDT**	**Frac. Dis.**
MGYP	Orphan	47.98	0.76	0.69	**65.17**	0.22	45.33	0.73
DP00070	IDR (100%)	71.83	0.92	0.77	**31.30**	**0.98**	39.60	0.02
DP03552	IDR (100%)	58.14	1.00	0.65	**51.60**	0.36	55.52	0.37

Frac. Dis. = fraction of residues with pLDDT <50; Rank = model ranking score.
Bold = appropriate confidence calibration for disordered proteins.

3.4 Computational Efficiency

Computational time varied dramatically across the three tools, with MSA-free models demonstrating substantial speed advantages over AlphaFold3 (Table 5).

Table 5. Computational time and speed ratios.

Protein	**Size**	**Runtime (s)**			**Ratio**	
	(aa)	**AF3**	**ESM**	**OF**	**AF3/E**	**AF3/O**
Monomers						
5L33	62	1042	4.22	5.05	247×	206×
1UBQ	76	1050	2.25	4.89	467×	215×
1QYS	106	977	2.66	5.27	367×	185×
Orphan and Disordered						
DP00070	140	2915	3.43	16.71	850×	174×
DP03552	352	1159	16.55	40.20	70×	29×
MGYP	421	993	17.46	70.16	57×	14×

Multi-chain excluded: AF3 models assemblies.

MSA-free models completed predictions in seconds across all protein types. ESMFold exhibited the fastest performance, ranging from 2.25 seconds (1UBQ, 76 residues) to 17.46 seconds (MGYP, 421 residues). OmegaFold required 4.64–70.16 seconds depending on protein length. In contrast, AlphaFold3 required 16–49 minutes per prediction (977–2915 seconds), representing a 57–850× slowdown compared to ESMFold. The speed differential was most pronounced for small proteins: ESMFold completed 1UBQ prediction 467× faster than AlphaFold3 (2.25 vs 1050 seconds). The maximum differential occurred for DP00070 (α-synuclein), where ESMFold required 3.43 seconds versus AlphaFold3's 2915 seconds an 850× difference.

The computational cost differential raises important practical considerations. While AlphaFold3 achieved superior accuracy on 4 of 5 monomers with experimental structures, RMSD differences between best and second-ranked models remained below 0.2 Å in all cases. For high-throughput applications or resource-constrained environments, ESMFold and OmegaFold offer compelling efficiency-accuracy trade-offs, delivering near-comparable accuracy at a fraction of computational cost.

4 Discussion

4.1 AlphaFold3 Maintains Critical MSA Dependence

Despite architectural modifications including replacement of Evoformer with Pairformer, AlphaFold3 has not eliminated dependence on multiple sequence alignments. This dependence persists but manifests differently across protein complexity levels.

For monomers, MSA dependence is moderate. AlphaFold3 achieved best performance on proteins with deep or shallow MSA (1UBQ, 1QYS) and maintained competitive accuracy even without MSA, winning 2 of 3 cases in this category (5L33, 6O35) with accuracy differences below 0.2 Å. This robustness suggests that evolutionary information is not critical for simple monomer structures, where AlphaFold3 can compensate through alternative information sources.

For multi-chain assemblies, MSA dependence becomes critical. AlphaFold3 failed catastrophically on 5 of 7 multi-chain proteins, producing RMSD 8–34 Å (2EKY: 26.067 Å, 6V67: 8.746 Å, 6O35: 13.645 Å, 6M0J: 27.765 Å, 7R6X: 34.098 Å). While most failures occurred on proteins with shallow or absent MSA, 7R6X demonstrates that even deep MSA does not guarantee success for complex hetero-oligomers. AlphaFold3 succeeded only on simpler

homo-oligomeric assemblies (1A3N: 0.480 Å, 6Z4U: 0.597 Å), suggesting that both MSA depth and assembly complexity determine prediction success. This clear performance dichotomy confirms that AlphaFold3 requires evolutionary co-variation signals for accurate quaternary structure modeling.

Analysis of failed predictions reveals an instructive paradox. When individual chains were extracted from failed multi-chain predictions, they exhibited accurate structures (6V67 chain A: 0.645 Å, 6O35 chain A: 0.424 Å, Section 3.1). AlphaFold3 thus correctly predicted monomer folding but failed to position subunits relative to each other. Without co-evolutionary signals indicating which residues from different chains likely interact, AlphaFold3 cannot reliably identify correct inter-chain interfaces. This failure mode (accurate monomer prediction but incorrect assembly) represents a fundamental limitation when MSAs lack sufficient depth to capture inter-chain co-evolution patterns.

4.2 AlphaFold3 Performance and Limitations

Systematic comparison across diverse protein types revealed that no single model is universally superior, with performance strongly dependent on protein characteristics and practical constraints.

AlphaFold3 strengths. For monomers with available MSA, AlphaFold3 demonstrated superior accuracy, achieving best predictions in 80% of cases (4 of 5 proteins) with RMSD typically below 0.5 Å. This performance confirms that AlphaFold3's architecture effectively leverages evolutionary information to produce high-quality structures when conditions are favorable. When MSA is available for monomer prediction, AlphaFold3 remains the tool offering highest absolute structural accuracy.

However, three critical limitations challenge AlphaFold3's position as the optimal choice in many practical contexts.

Limitation 1: Prohibitive computational cost. AlphaFold3 required 16–49 minutes per prediction, operating 57–850× slower than ESMFold depending on protein size. This computational overhead renders AlphaFold3 impractical for high-throughput screening applications requiring analysis of hundreds or thousands of protein variants. The cost-benefit analysis is particularly unfavorable: in 4 of 5 monomer cases, RMSD differences between AlphaFold3 and second-ranked models remained below 0.2 Å marginal gains that may not justify 57–850× longer computation times for many applications.

Limitation 2: Catastrophic failures on multi-chain assemblies without MSA. As detailed in Section 4.1, AlphaFold3's critical MSA dependence for multi-chain assemblies resulted in a 71% failure rate (5 of 7), limiting its reliability for novel protein complexes where MSA depth cannot be guaranteed.

Limitation 3: Systematic overconfidence. Confidence calibration issues emerged across multiple contexts. For multi-chain assemblies, catastrophic failures (RMSD 8–34 Å) occurred despite inappropriately high confidence scores (pLDDT 73–84, PTM 0.34–0.70). Notably, 7R6X exhibited pLDDT 80–82 despite RMSD 34.098 Å, though its low iPTM (0.25–0.31) partially detected the interface failure. This miscalibration suggests that AlphaFold3's confidence metrics do not reliably detect quaternary structure prediction failures.

For intrinsically disordered proteins, AlphaFold3 exhibited severe systematic overconfidence independent of MSA availability (Section 3.3). α-synuclein, experimentally validated as 100% disordered, received pLDDT 71.83 (a score indicating high-confidence well-folded structure) despite AlphaFold3's own disorder detection correctly identifying 92% disordered residues. This hallucination pattern, documented to affect 22% of experimentally validated disordered residues [5], represents a calibration issue distinct from MSA-related failures. The overconfidence persists regardless of MSA depth, indicating a fundamental limitation in AlphaFold3's confidence estimation for unstructured regions.

4.3 MSA-Free Models Demonstrate Superior Robustness

ESMFold and OmegaFold demonstrated remarkable robustness across diverse protein types, offering compelling alternatives to AlphaFold3 in many practical contexts.

ESMFold exceptional stability. ESMFold exhibited zero catastrophic failures across the entire dataset, with worst-case RMSD of 1.679 Å (6M0J) remaining within usable range for most applications. This consistent performance contrasts sharply with AlphaFold3's high variability, where predictions ranged from excellent (0.272 Å) to catastrophic failures (34.098 Å). ESMFold's robustness proved independent of both MSA depth and structural complexity, maintaining excellent performance

on proteins without MSA and achieving best accuracy on 4 of 7 multi-chain proteins for individual chain predictions (RMSD <1 Å in all cases, Section 3.2).

ESMFold's appropriate confidence calibration for intrinsically disordered proteins represents a critical advantage over AlphaFold3. For α-synuclein (100% disordered), ESMFold assigned appropriately low pLDDT (31.30) with 98% residues classified as disordered, accurately reflecting experimental validation. This contrasts with AlphaFold3's systematic overconfidence (pLDDT 71.83) for the same protein. Combined with 850× computational advantage over AlphaFold3 for this case, ESMFold's reliable disorder prediction makes it particularly suitable for proteome-wide disorder screening and metagenomic sequence analysis where structural state is uncertain.

For orphan proteins lacking sequence homologs, ESMFold demonstrated superior practical utility. On MGYP (metagenomic orphan, 421 residues), ESMFold produced moderate confidence (pLDDT 65.17) with 22% predicted disorder, suggesting partially structured regions. While AlphaFold3 and OmegaFold predicted predominantly disordered protein (pLDDT 47.98 and 45.33, disorder fractions 0.76 and 0.73), all three models agreed on absence of well-defined stable structure (consensus pLDDT <70). ESMFold's 57× speed advantage and demonstrated robustness without MSA make it optimal for rapid screening of metagenomic, synthetic, or de novo designed proteins.

OmegaFold competitive performance with calibration limitations. OmegaFold showed competitive accuracy for well-folded proteins, achieving best prediction on 2 proteins: 6M0J chain A (1.011 Å) and 7R6X chain D (0.308 Å), and maintaining reasonable performance across structured targets. However, analysis revealed significant confidence calibration issues for disordered proteins. For α-synuclein, OmegaFold assigned appropriate low pLDDT (39.60) but severely miscalibrated disorder fraction (2% versus 100% experimental validation), indicating fundamental limitations in disorder detection. This calibration failure limits confidence in OmegaFold's disorder-related predictions.

The 57–850× computational advantage of MSA-free models over AlphaFold3, combined with ESMFold's demonstrated robustness and appropriate confidence calibration, establishes MSA-free approaches as optimal choices for high-throughput applications, orphan protein analysis, disorder screening, and any context requiring guaranteed reliability without risk of catastrophic failure. AlphaFold3 retains advantages for maximum accuracy on well-characterized structured proteins with available MSA, but MSA-free models demonstrate superior practical utility across broader application contexts.

4.4 Practical Model Selection Guidelines

Systematic evaluation across diverse protein types enables evidence-based recommendations for model selection based on specific application requirements (Table 6).

Monomers with available MSA. AlphaFold3 remains the preferred choice when computational time is not limiting, achieving 80% win rate with exceptional accuracy (typically RMSD <0.5 Å). This recommendation applies when maximum structural accuracy is required for well-characterized proteins with sufficient sequence homologs.

Orphan proteins without sequence homologs. ESMFold is the optimal choice for proteins lacking detectable homologs, including metagenomic proteins, synthetic or de novo designed sequences. ESMFold demonstrated robustness without MSA, operating 57–850× faster than AlphaFold3, and exhibited zero catastrophic failures across the dataset. For metagenomic sequences of uncertain structural state (e.g., MGYP), ESMFold's appropriate calibration for both structured and disordered regions provides reliable confidence estimates (pLDDT 65.17 versus AlphaFold3 47.98, OmegaFold 45.33).

Intrinsically disordered proteins. ESMFold is strongly recommended for IDR analysis, producing appropriately low pLDDT scores with accurate disorder fractions matching standard interpretation scales (Section 3.3). AlphaFold3 should be avoided for IDR characterization due to systematic overconfidence, assigning high structural confidence to experimentally validated disordered proteins. OmegaFold exhibits severe disorder fraction miscalibration (2% versus 100% experimental validation for α-synuclein), limiting reliability. For comprehensive IDR characterization, predictions should be cross-validated with specialized disorder predictors (IUPred3, SPOT-Disorder) and, when feasible, experimental techniques (circular dichroism, SAXS).

Multi-chain assemblies. Recommendations depend on objectives. For individual chain struc-

Table 6. Model selection recommendations by application context.

Protein Type	Recommended	Justification
Monomer with MSA	AlphaFold3	Best accuracy (80% wins), MSA available
Orphan monomer	ESMFold	Robust without MSA, 57–850× faster, no failures
Disordered protein	ESMFold	Appropriate pLDDT calibration, avoids AF3 hallucinations and OmegaFold calibration issues
Multi-chain (monomer only)	ESMFold/Omega	High quality on individual chains (<1 Å), fast
Multi-chain (assembly)	AlphaFold3*	Only tool capable of assembly modeling, but 71% failure rate (5/7); complex hetero-oligomers challenging even with deep MSA → experimental validation necessary
Large protein (>1200 aa)	Domain splitting	GPU memory limit common to all three models
High-throughput	ESMFold	Optimal speed-accuracy trade-off

*Warning: Deep MSA improves but does not guarantee reliability for complex hetero-oligomers; confidence metrics unreliable.

tures only, ESMFold or OmegaFold provide excellent performance (typically RMSD <1 Å) with substantial speed advantages. For complete quaternary structure modeling, AlphaFold3 is the only tool capable of predicting intact assemblies. However, the 71% failure rate observed (5 of 7), with failures occurring even with deep MSA for complex hetero-oligomers (7R6X), necessitates experimental validation of predicted assemblies. AlphaFold3's confidence metrics (pLDDT, PTM, iPTM) do not reliably detect quaternary structure failures, requiring independent validation through biochemical or structural techniques.

High-throughput screening. For applications requiring analysis of hundreds or thousands of variants, ESMFold offers optimal speed-accuracy trade-off. Its computational efficiency enables large-scale analyses impractical with AlphaFold3 while maintaining excellent structural quality for most biological applications.

Very large proteins. Proteins exceeding approximately 1200 amino acids surpass current GPU memory constraints across all three tools. Domain-based approaches (predicting individual domains independently and assembling structures) may provide solutions for specific cases but require validation to confirm applicability.

4.5 Study Limitations

Several limitations constrain the scope and generalizability of this evaluation.

Dataset size and diversity. The dataset comprised 14 proteins, providing systematic coverage across key challenge categories (MSA depth variation, multi-chain assemblies, orphan proteins, intrinsically disordered proteins) but limiting statistical power for comprehensive performance characterization. Monomer evaluation included only 5 proteins with experimental structures, and multi-chain assessment covered 7 assemblies. Broader validation across larger protein sets, particularly expanded DisProt-validated disordered proteins and diverse multi-chain architectures, would strengthen confidence in observed performance patterns.

Hardware constraints. GPU memory limitations (NVIDIA A100, 80 GB VRAM) restricted evaluation to proteins below approximately 1200 amino acids. The SARS-CoV-2 Spike protein (6VXX, 1281 residues per chain) exceeded memory capacity across all three tools, with failures occurring during pairwise matrix generation (AlphaFold3) or language model processing (ESMFold, OmegaFold). This size limitation represents a shared challenge across current approaches rather than tool-specific constraints, but prevents assessment of model performance on large multi-domain proteins and macromolecular assemblies.

Experimental validation limitations. For proteins without experimental structures (MGYP, DP00070, DP03552), evaluation relied exclusively on computational confidence metrics rather than ground-truth structural data. While DisProt anno-

tations for DP00070 and DP03552 provide experimental disorder validation through circular dichroism, NMR spectroscopy, and SAXS, these proteins lack atomic-resolution reference structures. Assessment of MGYP's true structural state remains uncertain without experimental characterization. Predictions for these cases, particularly regarding disorder content, would benefit from experimental validation.

Confidence calibration analysis. Systematic confidence calibration assessment was limited to specific cases demonstrating clear miscalibration (AlphaFold3 overconfidence for multi-chain assemblies and disordered proteins, OmegaFold disorder fraction errors). Comprehensive calibration analysis across broader protein sets and confidence score ranges would provide more complete characterization of each tool's reliability indicators. The observed 22% hallucination rate for AlphaFold3 on disordered residues [5] suggests systematic issues warranting expanded investigation.

Scope limitations. Evaluation focused on protein structure prediction, excluding other AlphaFold3 capabilities (protein-ligand interactions, nucleic acid complexes, post-translational modifications). Performance patterns observed for protein-only predictions may not generalize to these alternative prediction tasks. Additionally, computational time measurements reflect specific hardware configuration and may vary across different infrastructure setups, though relative performance rankings should remain consistent.

Despite these limitations, systematic evaluation across diverse challenge categories (MSA depth variation, structural complexity, orphan sequences, and intrinsically disordered proteins) provides robust evidence for the performance boundaries and practical applicability of each tool.

5 Conclusion

This systematic evaluation of AlphaFold3, ESMFold, and OmegaFold across 14 carefully selected proteins provides evidence-based characterization of each tool's performance boundaries and practical applicability.

AlphaFold3 has not eliminated the MSA dependence that characterized AlphaFold2. While this dependence remains moderate for monomers (where AlphaFold3 achieved 80% best accuracy with RMSD typically below 0.5 Å) it becomes critical for multi-chain assemblies, resulting in 71% catastrophic failure rate (5 of 7), with failures occurring even with deep MSA for complex hetero-oligomers. Additionally, AlphaFold3 exhibits systematic overconfidence for intrinsically disordered proteins independent of MSA availability, assigning inappropriately high pLDDT scores (71.83 for 100% disordered α-synuclein) that generate structural hallucinations. These limitations, combined with 57–850× computational cost relative to MSA-free models, substantially constrain AlphaFold3's practical utility beyond well-characterized structured proteins with available MSA.

MSA-free models demonstrate superior robustness across diverse protein types. ESMFold exhibited zero catastrophic failures, appropriate confidence calibration for disordered proteins (pLDDT 31.30 for α-synuclein versus AlphaFold3's 71.83), and exceptional computational efficiency. These characteristics establish ESMFold as optimal choice for orphan protein analysis, disorder screening, and high-throughput applications. OmegaFold showed competitive performance for structured proteins but exhibited significant disorder calibration limitations.

Our findings enable evidence-based model selection: AlphaFold3 for maximum accuracy on well-characterized structured proteins with MSA; ESMFold for robustness, appropriate disorder calibration, and computational efficiency; OmegaFold for speed with competitive accuracy on structured targets. Critical recommendations include avoiding AlphaFold3 for intrinsically disordered protein analysis and requiring experimental validation for AlphaFold3 multi-chain predictions, particularly for complex hetero-oligomers where even deep MSA does not guarantee success.

Future work should expand evaluation to larger DisProt-validated datasets to comprehensively characterize confidence calibration patterns, assess performance on additional multi-chain architectures to refine MSA depth requirements, and validate predictions through experimental techniques such as circular dichroism and SAXS. As protein structure prediction tools continue evolving, systematic benchmarking across diverse challenge categories remains essential for defining practical performance boundaries and enabling optimal tool selection.

Acknowledgments

The authors thank the University of Malaya for providing access to computational resources. All predictions were performed on NVIDIA A100 GPUs (80 GB VRAM). TThis work was conducted as part of a four-month internship at the Faculty of Computer Science and Information Technology, University of Malaya, under the supervision of Anuar Badrul.

Data Availability

Predicted structures, confidence scores, and analysis scripts are available upon reasonable request from the corresponding author.

Supplementary Materials

The following supplementary files are available at the GitHub repository:

- **Detailed_Protein_Dataset.xlsx**: Complete protein dataset characteristics including PDB IDs, chain information, sequence lengths, and MSA depth classifications Available on GitHub
- **Detailed_Results_for_All_Models.xlsx**: Detailed prediction results for all three models including RMSD values, pLDDT scores, PTM/iPTM metrics, disorder fractions, and runtime measurements Available on GitHub

References

[1] Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2

[2] Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024). https://doi.org/10.1038/s41586-024-07487-w

[3] Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023). https://doi.org/10.1126/science.ade2574

[4] Wu, R. *et al.* High-resolution de novo structure prediction from primary sequence. *bioRxiv* (2022). https://doi.org/10.1101/2022.07.21.500999

[5] Gopalan, S. & Narayanan, S. Hallucinations in AlphaFold3 for Intrinsically Disordered Proteins with disorder in Biological Process Residues. *arXiv preprint* arXiv:2510.15939 (2024). https://doi.org/10.48550/arXiv.2510.15939

[6] Mylonas, E. *et al.* Domain conformation of tau protein studied by solution small-angle X-ray scattering. *Biochemistry* **47**, 10345–10353 (2008). https://doi.org/10.1021/bi800900d

[7] Aspromonte, M. C. *et al.* DisProt in 2024: improving function annotation of intrinsically disordered proteins. *Nucleic Acids Research* **52**, D434–D441 (2024). https://doi.org/10.1093/nar/gkad928