Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Parallel Bias Is Not Universal in Deep Learning Co-Folding Models: A Pre-Registered Benchmark of Amyloid Sheet Architecture

Giriish Purushothaman

giriish2009@gmail.com

Bridgeland High School

Research Article

Keywords: AlphaFold 3, amyloid, amyloid-beta, cross-beta sheet, parallel and antiparallel topology, protein structure prediction, co-folding models, benchmarking

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11861249/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Parallel Bias Is Not Universal in Deep Learning Co-Folding Models: A Pre-Registered Benchmark of Amyloid Sheet Architecture

Giriish Purushothaman

Abstract

Structural polymorphism in amyloid fibrils often diverges at the level of sheet topology, determined by whether constituent β-strands pack parallel in-register or antiparallel. While biomolecular co-folding networks such as AlphaFold 3 (AF3) are increasingly deployed to model macromolecular interfaces and mutational effects, established evaluation metrics such as TM-score and interface confidence (ipTM) assess global envelope packing without encoding chain polarity. Consequently, whether these models distinguish quaternary sheet orientation or collapse to an inductive architectural preference remains unexamined.

Here, we present `sheetarch`, a fail-closed geometric descriptor suite validated across 15 experimental fibril structures to quantify inter-chain hydrogen-bond networks, tangent vector alignments, and segment-level register. In a retrospective survey of 160 jobs (800 models) across full-length amyloid-β (Aβ), AF3 exclusively predicted parallel fibrils, missing the known metastable antiparallel polymorph of $A\beta_{1-40}$ D23N (Iowa). A pre-registered confirmatory benchmark (Phase 3; 35 jobs, 175 models across seven segments) refuted the hypothesis of a general parallel bias: AF3 accurately assembled antiparallel steric-zipper sheets for insulin segments (system-mean cosines below -0.90) and parallel controls, while hedging on the Aβ segment MVGGVV.

A second pre-registered benchmark (P6; 30 jobs, 150 models) prospectively tested this apparent sequence-architecture conjunction across additional Aβ segments. P6 produced the pre-registered outcome of exactly one correct target: AF3 correctly assembled KLVFFA as antiparallel (5 of 5 jobs) but hedged on GAIIGL (2 of 5 jobs), while parallel controls maintained fidelity. Across exploratory evaluations, alternative architectures displayed divergent behaviors: Boltz-1 reinforced parallel packing, Chai-1 exhibited extensive topological disagreement, and Boltz-2 favored negative cosines without recovering clean fibrillar consensus.

Across retrospective Aβ assemblies, partial antiparallel geometry concentrated systematically in lower-ranked, lower-confidence samples (24 of 160 jobs overall versus 1 of 160 top-ranked models; paired ipTM difference -0.059, $p = 6.3 \times 10^{-5}$). These findings demonstrate that AF3 is not architecture-blind, but its fidelity narrows across specific Aβ targets. Researchers should score complete diffusion ensembles rather than top-ranked models alone when assessing polymorphic fibrillar assemblies.

1. Introduction

Amyloid fibrils are ordered protein aggregates characterized by cross-β architecture, in which individual polypeptide chains stack perpendicular to the fibril axis into continuous β-sheets. A central feature of amyloid biology is structural polymorphism: an identical amino acid sequence can assemble into structurally distinct fibrillar folds (polymorphs) that exhibit divergent biological properties (Sawaya et al., 2007; Tycko, 2015). At the most elementary level of quaternary organization, polymorphs diverge in strand orientation, determined by whether adjacent hydrogen-bonded chains in a cross-β sheet run parallel in-register or antiparallel. A prominent biological example is the familial Iowa mutation of amyloid-β ($A\beta_{1-40}$ D23N), for which both antiparallel (PDB 2LNQ) and parallel in-register (PDB 2MPZ) fibril polymorphs have been experimentally determined; the antiparallel assembly is metastable relative to the parallel form and gives way to it over repeated seeding (Qiang et al., 2012).

With advances in deep learning architectures capable of predicting macromolecular structures, such as AlphaFold 2 and AlphaFold 3 (AF3; Jumper et al., 2021; Abramson et al., 2024), co-folding models are increasingly used to model macromolecular interfaces and assess the structural consequences of disease-associated mutations. By design, however, AF3 generates individual structural models via a diffusion process accompanied by predicted confidence scores, rather than enumerating conformational ensembles across complex, multi-state free-energy landscapes. While alternative conformations can sometimes be accessed via multiple sequence alignment (MSA) subsampling or clustering (del Álamo et al., 2022; Wayment-Steele et al., 2024), these methods were developed for deep monomeric alignments rather than quaternary amyloid sheets.

Standard structural assessment metrics often fail to evaluate whether a deep learning framework accurately models sheet orientation. Commonly used metrics, such as per-chain TM-score, pLDDT, and interface confidence (ipTM), reward overall shape alignment, local stereochemical quality, and contact density, but do not encode chain polarity. Consequently, a predicted model can achieve favorable structural metrics even when the constituent β-sheets run in the wrong direction. To our knowledge, no systematic benchmark has addressed the direct architectural question: when tasked with assembling multi-chain amyloid complexes, does a model reliably distinguish between parallel and antiparallel topology, or does it collapse toward a default inductive orientation? Establishing this distinction is geometrically non-trivial, as automated descriptors must handle irregular inter-chain hydrogen-bond networks, boundary effects, and lattice symmetries without introducing systematic classification artifacts.

This study investigates whether AlphaFold 3 correctly determines amyloid β-sheet architecture or defaults to an architectural bias. We developed `sheetarch`, an automated, fail-closed classification package that

evaluates sheet topology using three geometric metrics calibrated against a reference suite of 15 experimentally determined fibrillar structures. We combined retrospective evaluations across full-length Aβ assemblies (160 jobs, 800 models) with two prospectively committed, pre-registered benchmarks (Phase 3 and P6) whose decision boundaries and interpretive consequences were frozen before executing prediction jobs. The study followed an empirical trajectory: initial retrospective findings suggested that AF3 suffered from a universal inductive bias toward parallel sheets. Our first pre-registered test (Phase 3) refuted this hypothesis by demonstrating that AF3 correctly constructs antiparallel sheets for non-Aβ segments from insulin. Observing that failures remained localized to Aβ-derived sequences with antiparallel ground truth, we conducted a second pre-registered test (P6), which further narrowed the claim: AF3 succeeds on some antiparallel Aβ segments (KLVFFA) while hedging or failing on others (MVGGVV, GAIIGL). Because each confirmatory phase was bound to pre-registered decision rules, empirical outcomes directly constrained and narrowed the theoretical claims.

The remainder of this report is organized as follows. Section 2 (Methods) describes the prediction protocols, the experimental ground-truth reference suite, the mathematical implementation of the `sheetarch` descriptors, and the pre-registered statistical decision frameworks. Section 3 (Results) details the retrospective Aβ arms, the Phase 3 refutation, the P6 replication and boundary narrowing, and rank-dependent metric distributions. Section 4 (Discussion and Limitations) contextualizes these architectural constraints, examines the implications for interpreting mutational landscapes, and outlines methodological limitations. Section 5 concludes the report.

2. Methods

2.1 Overview and Study Design

This study evaluated whether AlphaFold 3 (AF3) accurately predicts the relative strand orientation (parallel versus antiparallel cross-β architecture) of amyloid assemblies across distinct sequence and polymorph regimes. The experimental design comprised three parts: a retrospective exploration across full-length amyloid-β (Aβ) assemblies, two pre-registered confirmatory benchmarks (Phase 3 and P6) with decision boundaries committed prior to job submission, and an experimental reference suite to calibrate the automated geometric classification package `sheetarch` (Table 1). The 25-job Aβ42 four-chain arm listed in Table 1 was generated for a previously accepted conference study (Purushothaman, 2026) and is re-analysed here under the sheetarch descriptors; the architectural analysis reported for that arm is new to this work, and all other arms were generated for the present study.

Table 1. Summary of computational prediction arms, structural systems, and sampling scale.

Arm	Status	Systems / Targets	Jobs	Models	Sampling Parameters
BIBM Aβ42 4-Chain	Exploratory	Aβ42 WT, E22G, E22Q, D23N, F19P	25	125	5 seeds, 5 samples/job
Aβ40/42 Hexamer	Exploratory	35 DMS variants + WT Aβ42; 3 Aβ40 variants × 4 template conditions	135	675	DMS variants: 2 seeds; WT & Aβ40: 5 seeds; 5 samples/job
Phase 3 Benchmark	Confirmatory	7 short segments across 5 proteins (insulin, tau, prion, Sup35, Aβ)	35	175	5 seeds, 5 samples/job
P6 Benchmark (Arms A & B)	Confirmatory	2 Aβ antiparallel targets (KLVFFA, GAIIGL) + 4 Aβ parallel controls	30	150	5 seeds, 5 samples/job
P6 Benchmark (Arm C)	Exploratory	MVGGVVIA (dual-polymorph Aβ segment)	5	25	5 seeds, 5 samples/job
Total[1]			**230**	**1,150**	

2.2 Structure Prediction

All primary structural predictions were generated using AlphaFold 3 via the AlphaFold Server web interface (Abramson et al., 2024) between June and November 2026. Standard server default settings were applied across all runs, with structural template usage disabled except in designated template-condition arms. Every system was sampled across assigned seeds at 5 diffusion samples per job.

Random seeds were fixed deterministically and recorded prior to submission:

- Phase 3 used `SEED_RNG = 20261017`.
- P6 used `SEED_RNG = 20261102`.
- The Aβ40 hexamer, deep mutational scanning (DMS), and cross-model batches derived seeds by applying a SHA-256 hash to the job name.

[1] Total counts include the 5 exploratory jobs (25 models) in Arm C.

In all cases, seeds were written into the batch configuration files before submission. Target sequences and chain counts were extracted from the deposited atomic coordinates (mmCIF/PDB format) rather than literature text or figures, avoiding discrepancies between published figures and deposited atomic coordinates.

The Aβ40 template-condition arm evaluated four conditions using exact date cutoffs:

1. *No template*: Template retrieval disabled.
2. *2021-09-30 cutoff*: `maxTemplateDate` set to 2021-09-30, matching AlphaFold 3's primary training cutoff date.
3. *2025-02-03 cutoff*: `maxTemplateDate` set to 2025-02-03, allowing retrieval of more recent structural depositions.
4. *Forced antiparallel template*: The antiparallel Aβ40 D23N coordinate structure (PDB 2LNQ) supplied directly as the structural template.

Prior to scoring, all returned outputs were verified against original submission manifests for target sequence identity, chain count, template flags, and initial random seeds. Across all 30 jobs in the P6 benchmark, zero configuration mismatches were detected. Model indices assigned by the extraction pipeline mapped identically to AlphaFold 3 internal confidence rankings (ordered by `ranking_score`) across 225 of 225 verified jobs.

2.3 Experimental Ground Truth

The experimental reference suite comprised 15 experimentally determined amyloid structures (11 parallel in-register and 4 antiparallel β-sheet architectures) determined by solid-state nuclear magnetic resonance (ssNMR), cryogenic electron microscopy (cryo-EM), and X-ray microcrystallography (Table 2).

A reference pair is formed by PDB entries 2LNQ (antiparallel, full-length 40-residue sequence) and 2MPZ (parallel, construct spanning residues 15–40), both carrying the Iowa mutation (D23N). This matched sequence core provides an experimental control that separates architecture from sequence.

For microcrystalline steric-zipper structures determined by X-ray crystallography, the asymmetric unit contains one peptide chain per β-sheet, with inter-chain sheet packing generated by space-group lattice symmetry. Unexpanded coordinate files show no inter-chain sheet contacts. Coordinates were therefore processed through an assembly expansion module using the symmetry operations declared in the coordinate metadata (REMARK 350 / `_pdbx_struct_assembly_gen`). Ground truth classifications for all P6 short segments were derived from atomic coordinates and checked against the source papers (Colletier et al., 2011; Landau et al., 2011), with 9 of 9 agreeing.

Table 2. Experimental reference structures utilized for calibration and geometric validation.

PDB ID	Construct / Segment	Source Protein	Experimental Method	Ground Truth Architecture
2LMN	Full-length $A\beta_{1-40}$ WT	Amyloid-β	ssNMR	Parallel
2LMP	Full-length $A\beta_{1-40}$ WT	Amyloid-β	ssNMR	Parallel
2M4J	Full-length $A\beta_{1-40}$ WT	Amyloid-β	ssNMR	Parallel
6SHS	Full-length $A\beta_{1-40}$	Amyloid-β	Cryo-EM	Parallel
8J47	Full-length $A\beta_{1-40}$ E22G (Arctic)	Amyloid-β	Cryo-EM	Parallel
5OQV	Full-length $A\beta_{1-42}$ WT	Amyloid-β	Cryo-EM	Parallel
2MPZ	$A\beta_{15-40}$ D23N (Iowa)	Amyloid-β	ssNMR + EM + Rosetta	Parallel
2ON9	VQIVYK (306–311)	Tau	X-ray crystallography	Parallel
2OL9	SNQNNF (170–175)	Prion	X-ray crystallography	Parallel
2OMM	GNNQQNY (7–13)	Sup35	X-ray crystallography	Parallel
2ONV	GGVVIA (37–42)	Amyloid-β	X-ray crystallography	Parallel
2LNQ	Full-length $A\beta_{1-40}$ D23N (Iowa)	Amyloid-β	ssNMR	Antiparallel
2OMQ	VEALYL (B12–B17)	Insulin	X-ray crystallography	Antiparallel
2OMP	LYQLEN (A13–A18)	Insulin	X-ray crystallography	Antiparallel
2ONA	MVGGVV (35–40)	Amyloid-β	X-ray crystallography	Antiparallel

2.4 Architecture Descriptors

Amyloid architecture was defined by whether neighboring, hydrogen-bonded chains in the cross-β sheet run parallel or antiparallel. The `sheetarch` package quantifies sheet topology using three descriptors.

Bridges Descriptor. Inter-chain hydrogen bonding networks were identified using the Kabsch and Sander (1983) electrostatic definition in DSSP across backbone N, C, O atoms, with hydrogen positions computed geometrically. Adjacent residue pairs form a β-bridge when electrostatic interaction energy satisfies E < -0.5 kcal/mol. Bridges were classified based on bonding connectivity: an antiparallel bridge requires reciprocal bonds between residues i and j, or paired bonds linking i-1 → j+1 and j-1 → i+1. A parallel bridge requires bonds linking i-1 → j and j → i+1, or j-1 → i and i → j+1.

Direction Descriptor. Local chain orientation was computed from backbone directional tangent vectors between spatial $C\alpha_{i-1} \rightarrow C\alpha_{i+1}$ coordinates of spatially paired residues. Contacting chain pairs required ≥ 6 inter-chain residue pairs with pairwise Cα–Cα < 6.0 Å. For each aligned pair, the local directional cosine was calculated as:

$$\cos\theta = (v_A \cdot v_B)/(\|v_A\| \|v_B\|)$$

where v_A and v_B represent local segment tangent vectors on neighboring chains, ranging from +1.0 (parallel) to -1.0 (antiparallel). The fraction of contacting pairs with $\cos\theta < 0$ determines the categorical direction call, while the mean tangent cosine across contacting chain pairs serves as the quantitative endpoint for statistical tests.

Register Descriptor. Inter-chain residue register was determined by calculating the Spearman rank-order correlation coefficient (ρ) of paired residue indices along contacting interfaces within contiguous β-segments (≥ 4 residues) as an internal consistency check.

Consensus Rules. Models were classified as parallel or antiparallel if Bridges and Direction agreed. Differing calls were assigned as `DISAGREEMENT`, absence of contacts as `indeterminate`, and concordant intermediate states (0.10 ≤ antiparallel fraction ≤ 0.90) as `mixed`. The metric implementation follows three rules: fail closed to `indeterminate` or `DISAGREEMENT`, keep mixed outcomes, and never pool rank statistics across discontinuous segments.

2.5 Validation

Validation of `sheetarch` was conducted across the 15 experimental reference structures. All 15 structures (11 parallel, 4 antiparallel) were classified correctly across all three descriptors. Because this reference suite was used to identify edge cases and calibrate the descriptors during development, these numbers reflect development performance rather than held-out prospective validation.

Evaluating the suite identified an artifact in the Register metric: pooling residue pairs across both β-segments of the hairpin fold (residues 15–24 and 29–40 in PDB 2LNQ) produced an apparent correlation of $\rho = +0.429$, whereas segment-confined computation correctly yielded $\rho = -0.988$. Restricting correlation calculations to contiguous segments corrected this behavior and established the segment-level evaluation rule.

2.6 Statistical Analysis

The primary unit of statistical analysis was the independent prediction job (N = 5 random seeds per target, except where noted), averaging the 5 diffusion samples within each job to yield a mean job cosine ($\bar{c}_{job}$). Any job with a non-finite sample cosine was scored as indeterminate while remaining in the denominator.

Pre-registered decision rules were evaluated across the 5 job means:

- *Phase 3 Refutation Criteria (Parallel Bias):* Hypothesis rejected if $\bar{c}_{job} < 0.0$ in ≥ 3 of 5 jobs for non-Aβ antiparallel systems.
- *P6 Antiparallel Target Success:* Correct antiparallel recovery defined as $\bar{c}_{job} < 0.0$ in ≥ 3 of 5 jobs.
- *P6 Parallel Control Success:* Correct parallel recovery defined as $\bar{c}_{job} > +0.50$ in ≥ 4 of 5 jobs.

Proportion confidence intervals were calculated using the exact Clopper–Pearson method (Clopper & Pearson, 1934). Equivalence across template conditions was tested using Welch's TOST (Schuirmann, 1987) with a pre-specified margin of $\Delta = \pm 0.10$ cosine units (5.7% of the 1.766 gap between the parallel and antiparallel reference sets, fixed before inspecting the predictions). Sequence variation was characterized by estimating between-sequence standard deviation (σ_{seq}) via a linear random-effects model with bootstrapped 95% upper bounds. Rank-order concentration within jobs was tested using an exact job-conditioned rank test, and confidence differences (`ipTM`) via paired Wilcoxon signed-rank and Mann–Whitney U tests. Analyses used Python v3.10.12, NumPy v2.2.6, SciPy v1.15.3, and Gemmi v0.7.5.

2.7 Cross-Model Comparison

To evaluate whether observed topological behaviors were unique to AlphaFold 3 or shared across modern biomolecular deep-learning architectures, comparative predictions were performed using Boltz-1 (Wohlwend et al., 2024), Boltz-2 (Passaro et al., 2025), and Chai-1 (Chai Discovery, 2024). Multiple sequence alignments were generated using the ColabFold pipeline (Mirdita et al., 2022).

Each benchmark evaluated hexameric assemblies of three targets: $A\beta_{1-40}$ wild-type, D23N (Iowa), and E22G (Arctic). Modeling was configured across 5 pre-specified seeds per system, sampling 5 structural models per seed without structural templates, with seeds fixed in advance using SHA-256 hashes of the job names. Computational jobs were run in isolated virtual environments on Kaggle-hosted NVIDIA Tesla T4 compute instances running Python 3.11. Generated structural coordinates were downloaded and evaluated locally using the identical `sheetarch` pipeline.

2.8 Reproducibility

All analysis scripts, frozen manifests, pre-registration protocols, and the experimental reference structures are tracked in the study repository and are available from the author upon request. Phase 3 reproduces from raw prediction outputs. Every pre-registration commit, including formal protocol amendments, preceded job submission and data generation; all repository commit hashes and UTC timestamps are permanently logged in version control.

3. Results

3.1 In the Aβ Arms, AF3 Never Builds an Antiparallel Fibril

Across both retrospective arms, AlphaFold 3 (AF3) never predicted an antiparallel consensus β-sheet architecture for full-length amyloid-β (Aβ). Across 160 independent prediction jobs comprising 800 structural models (5 diffusion samples per job), zero models reached an antiparallel consensus call (0/800, exact 95% binomial confidence interval [0.000, 0.0046] at the uncorrected model level; Table 3). Because individual diffusion samples within a job share an identical multiple sequence alignment and random seed, the appropriate unit of analysis is the independent job: at the job level, 0 of 160 jobs produced an antiparallel consensus architecture, placing an exact 95% upper bound on the incidence rate below 0.023 ([0.000, 0.0228], Clopper–Pearson).

This absence of antiparallel predictions held across all 36 tested $A\beta_{1-42}$ sequence variants, all 3 $A\beta_{1-40}$ variants, and across all 4 template conditions. In every job, the mean local backbone directional cosine ($\bar{c}_{job}$) ranged from +0.763 to +1.000, sitting 1.573 to 1.810 cosine units away from the deposited antiparallel reference structure (c = -0.810).

Table 3. Consensus architectural classification across all 160 retrospective Aβ prediction jobs (800 models).

Arm	Jobs	Models	Parallel	Antiparallel	Mixed	DISAGREEMENT	Indeterminate	Job-Mean Cosine Range ($\bar{c}_{job}$)
BIBM Aβ42 (4-chain)	25	125	125	0	0	0	0	+0.987 to +1.000
Aβ42 DMS (6-chain)	75	375	357	0	10	8	0	+0.763 to +1.000
Aβ40 WT & Mutants	60	300	287	0	5	8	0	+0.766 to +1.000
Total	**160**	**800**	**769**	**0**	**15**	**16**	**0**	**+0.763 to +1.000**

A key test within this retrospective cohort is the $A\beta_{1-40}$ D23N (Iowa) mutation. Experimentally, D23N forms both parallel and antiparallel fibrils under varying fibrillization conditions: PDB 2LNQ is an antiparallel fibril determined by solid-state NMR (cosine -0.810; 82 antiparallel β-bridges, 0 parallel β-bridges), while PDB 2MPZ is a parallel fibril (+0.979; 360 parallel β-bridges, 0 antiparallel β-bridges).

Across 20 jobs (100 models) spanning 4 template conditions, AF3 generated 96 cleanly parallel models, 2 mixed models, and 2 descriptor disagreements; zero models were classified as antiparallel. Forcing the monomer coordinates of 2LNQ itself as the structural template did not produce an antiparallel fibril (0/25 models, 0/5 jobs).

This outcome is a coverage failure on a polymorphic energy landscape: AF3 reliably predicts one physically valid polymorph (the parallel form) while failing to sample a known alternative metastable polymorph (Qiang et al., 2012). This finding describes the full-length Aβ arms specifically; it does not constitute a general claim regarding AF3's capacity to predict antiparallel architecture, which is directly addressed in Section 3.5.

Neither primary sequence variation nor structural templates altered this outcome. Across the 36-variant $A\beta_{1-42}$ panel, the between-sequence standard deviation was $\sigma_{seq} = 0.017$ cosine units, and all pairwise equivalence contrasts across the four template conditions fell within the pre-specified ± 0.10 equivalence margin. Supplying the antiparallel 2LNQ monomer as a structural template failed to shift predicted assemblies toward antiparallel packing.

3.2 Partial Antiparallel Content Hides in Low-Ranked Samples

Although full antiparallel fibrils were absent, sub-fibrillar antiparallel geometry was not entirely unrepresented. To quantify partial antiparallel content without arbitrary thresholding, models were scored by the presence of at least one antiparallel backbone β-bridge (DSSP electrostatic energy E < -0.5 kcal/mol).

At least one antiparallel β-bridge was detected in 24 of 160 jobs (15.0%, exact 95% CI [0.099, 0.215]), comprising 29 of 800 individual models (3.6%, [0.024, 0.052]). When evaluating only the top-ranked model per job (Rank 1, selected by AF3's `ranking_score`), an antiparallel bridge was present in only 1 of 160 jobs (0.6%, [0.000, 0.034]). Across confidence ranks 1 through 5, the model count displaying an antiparallel bridge increased monotonically (Figure 1).

An exact job-conditioned rank test demonstrated significant concentration of antiparallel bridges in lower-ranked diffusion samples, yielding an observed within-job rank sum of 89 against an expected null rank sum of 58.0 ($p = 9.1 \times 10^{-6}$). In the supporting test on the 25 jobs displaying mixed architecture, the mean rank of antiparallel-bearing models was 4.0 compared to an expected mean rank of 3.0 under uniform sampling (Wilcoxon signed-rank test, $p = 5.3 \times 10^{-4}$).

As a sensitivity check, the continuous "not cleanly parallel" metric (antiparallel direction fraction > 0.10) flagged 25 of 160 jobs (15.6%, job-conditioned rank test $p = 6.7 \times 10^{-5}$). This directional metric was stable across threshold variations from 0.05 to 0.15, but the signal was lost at a threshold of 0.20. Because

the calibrated metric exhibits threshold dependency, the parameter-free bridge criterion carries the formal conclusion.

3.3 AF3 Assigns Those Samples Lower Confidence

Within the 24 jobs containing partial antiparallel geometry, AF3 assigned lower internal confidence scores to the models displaying antiparallel β-bridges. Evaluating paired differences within each job (comparing the mean score of bridge-positive samples against the mean score of bridge-negative samples within that same job), the interface confidence metric (`ipTM`) was systematically lower in the presence of antiparallel bridges: mean ipTM was 0.254 for bridge-bearing models versus 0.314 for bridge-negative models (paired mean difference Δ = -0.059; paired Wilcoxon signed-rank test, $p = 6.3 \times 10^{-5}$; Table 4). In 20 of the 24 jobs, the bridge-bearing sample received a lower ipTM score than its intra-job counterparts (two-sided binomial sign test, $p = 1.5 \times 10^{-3}$).

Table 4. Paired comparison of AF3 internal quality metrics between bridge-positive and bridge-negative samples within the 24 jobs exhibiting partial antiparallel content.

Metric	Bridge-Positive Samples (Mean)	Bridge-Negative Samples (Mean)	Mean Paired Difference (Δ)	Wilcoxon p-value
ipTM	0.254	0.314	-0.059	$p = 6.3 \times 10^{-5}$
pTM	0.291	0.342	-0.051	$p = 1.7 \times 10^{-4}$
pLDDT	56.12	59.35	-3.23	$p = 1.75 \times 10^{-4}$
ranking_score	0.762	0.820	-0.057	$p = 6.3 \times 10^{-5}$

Global confidence scores, however, offer limited discriminative utility for identifying native fibrillar packing: across all 800 models, ipTM values spanned a low overall range from 0.160 to 0.650. Zero of 800 models achieved an ipTM of 0.80, and only 4 of 800 models reached 0.60. In the specific case of $A\beta_{1-40}$ D23N, the diffusion samples displaying directional alignment with the deposited antiparallel polymorph are assigned lower confidence scores by AF3's internal ranking function. The rank concentration and paired confidence score distributions are illustrated in Figure 1.

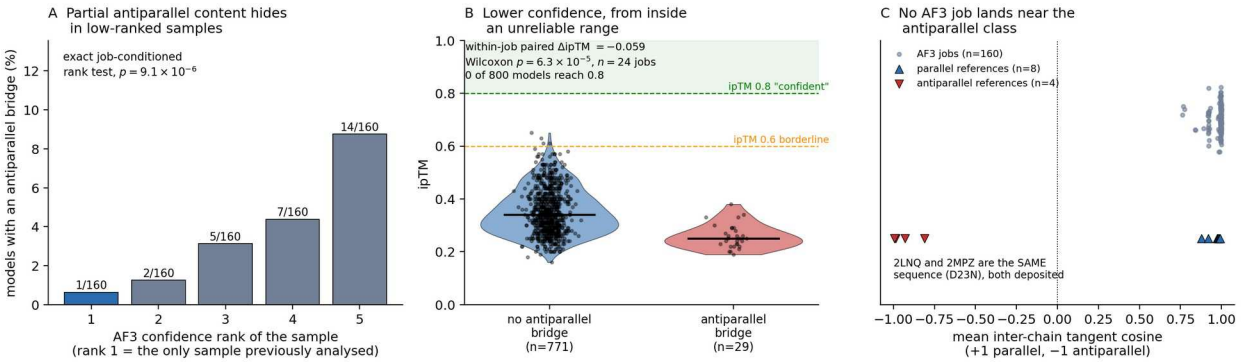

Figure 1. Distribution of partial antiparallel geometry and internal confidence scores across retrospective Aβ assemblies.

(A) Percentage of models containing at least one antiparallel β-bridge across AF3 internal confidence ranks (Ranks 1 to 5, where Rank 1 represents the top-ranked model; n = 160 jobs, 800 models). Antiparallel bridge incidence increases monotonically from Rank 1 (1 of 160) to Rank 5 (14 of 160; exact job-conditioned rank test, $p = 9.1 \times 10^{-6}$).

(B) Distribution of interface predicted TM-scores (ipTM) split between models lacking antiparallel β-bridges (blue violin; n = 771) and models containing at least one antiparallel β-bridge (red violin; n = 29). Solid horizontal bars mark sample medians. Dashed green and orange lines denote published AF3 confidence thresholds (ipTM 0.80 and 0.60, respectively). Within the 24 jobs containing bridge-positive samples, paired ipTM was significantly lower for bridge-positive models (mean difference -0.059; paired Wilcoxon signed-rank test, $p = 6.3 \times 10^{-5}$).

(C) Mean inter-chain tangent cosine across all 160 retrospective AF3 jobs (circles; n = 5 seeds per system) plotted alongside experimental ground-truth reference structures (triangles). Upward-pointing triangles mark experimental parallel reference structures plotted in this panel (n = 8); downward-pointing triangles mark experimental antiparallel reference structures (n = 4), including the matched-sequence Iowa mutant pair PDB 2LNQ (antiparallel) and 2MPZ (parallel). Vertical line at 0.0 separates parallel (> 0.0) from antiparallel (< 0.0) directional alignment.

Source: Generated by arch_figure.py from architecture_all_models.csv (160 jobs, 800 models). Figure created by the author.

3.4 Phase 3: Pre-Registered Test Refutes a General Parallel Bias

Phase 3 evaluated the hypothesis that AF3 exhibits a global inductive bias toward parallel sheets across all amyloid-forming sequences. All 35 prediction jobs (175 models across 7 peptide segments from 5 proteins) executed with zero indeterminate calls and were evaluated against pre-registered decision rules (Table 5).

The prediction that antiparallel systems would be forced parallel failed: zero jobs yielded $\bar{c}_{job} > +0.50$ for the antiparallel targets. Instead, the formal refutation rule fired: both non-Aβ antiparallel segments (insulin VEALYL and LYQLEN) yielded 5 of 5 job means below 0.0 ($\bar{c}$ = -0.980 and -0.940, respectively), correctly recovering native antiparallel architecture. All four parallel control segments (tau VQIVYK, prion SNQNNF, Sup35 GNNQQNY, and Aβ GGVVIA) produced 5 of 5 job means above +0.50. These outcomes rule out three alternative explanations: that AF3 cannot predict antiparallel architecture (ruled out by insulin), that AF3 forces all Aβ-derived segments parallel (ruled out by

GGVVIA), and that short segment length impedes prediction (ruled out by GGVVIA at the identical hexapeptide length).

The sole failing segment was the Aβ hexapeptide MVGGVV (PDB 2ONA; true architecture antiparallel), which produced a hedged mean cosine of +0.253. In a permutation test permuting AF3's internal confidence ranks within each job, antiparallel character concentrated significantly at lower confidence ranks ($p = 0.0095$). Under the calibrated direction cutoff (0.90), 2 of 25 diffusion samples recovered the correct full antiparallel architecture, both at Rank 5; across 29 alternative cutoffs, 16 yielded $p > 0.05$. A binomial resampling model indicates that drawing 28 diffusion samples (~ 6 jobs) yields a 90% probability of observing at least one fully antiparallel MVGGVV model, whereas for full-length $A\beta_{1-40}$ D23N, zero antiparallel models appeared across 20 jobs (rate < 0.168).

ColabFold MMseqs2 queries confirmed that all 7 Phase 3 segments exhibited an effective multiple sequence alignment (MSA) depth of 1, demonstrating that antiparallel sheets were assembled without coevolutionary covariation signals. Job means and the rank-dependent distribution of MVGGVV are plotted in Figure 2. Across the panel, AF3 correctly resolved 6 of 7 segments (exact 95% CI [0.421, 0.996]).

Table 5. Confirmatory Phase 3 benchmark results across 7 amyloid segments (35 jobs, 175 models).

Target Segment	Source Protein	Reference PDB	True Architecture	AF3 Mean Cosine ($\bar{c}$)	Job Outcome (Means < 0.0)	Pre-Registered Call
VEALYL	Insulin	2OMQ	Antiparallel	-0.980	5 of 5	Antiparallel (Correct)
LYQLEN	Insulin	2OMP	Antiparallel	-0.940	5 of 5	Antiparallel (Correct)
MVGGVV	Amyloid-β	2ONA	Antiparallel	+0.253	0 of 5	Hedged (Incorrect)
VQIVYK	Tau	2ON9	Parallel	+1.000	0 of 5	Parallel (Correct)
SNQNNF	Prion	2OL9	Parallel	+1.000	0 of 5	Parallel (Correct)
GNNQQNY	Sup35	2OMM	Parallel	+0.985	0 of 5	Parallel (Correct)
GGVVIA	Amyloid-β	2ONV	Parallel	+1.000	0 of 5	Parallel (Correct)

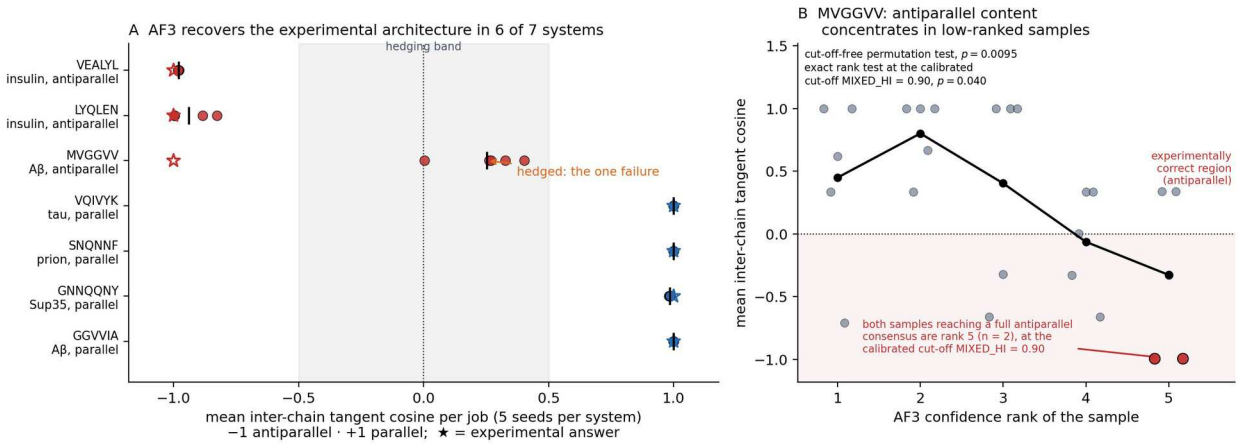

Figure 2. Confirmatory Phase 3 benchmark across seven amyloid segments.

(A) Mean inter-chain tangent cosine per job (circles; n = 5 seeds per segment) across seven target segments from five amyloidogenic proteins. Open stars indicate experimentally determined ground-truth values. The shaded vertical region between -0.50 and +0.50 marks the intermediate hedging zone. AF3 recovered native antiparallel architecture for both insulin controls (VEALYL and LYQLEN) and native parallel architecture for all four parallel controls, while hedging on Aβ MVGGVV (system mean +0.253).

(B) Distribution of inter-chain tangent cosines for all 25 individual diffusion samples of Aβ MVGGVV plotted against AF3 confidence rank (Ranks 1 to 5). Black line connects mean cosine values across ranks. Shaded horizontal band denotes the experimentally correct antiparallel region (cosine < 0.0). Full antiparallel consensus calls (under the calibrated MIXED_HI = 0.90 cutoff) appear exclusively at Rank 5 (lowest confidence rank; cutoff-free permutation test, p = 0.0095).

Source: Generated by phase3_figure.py from phase3_architecture.csv (35 jobs, 175 models). Figure created by the author.

3.5 P6: Replicating the Failing Cell Narrows the Claim

Because the Phase 3 failure coincided with an Aβ sequence context and an antiparallel ground truth, the P6 benchmark evaluated this conjunction prospectively under frozen decision rules across two additional Aβ antiparallel targets and four parallel controls (30 jobs, 150 models; Table 6).

Table 6. Confirmatory P6 benchmark results across 6 Aβ peptide segments (30 jobs, 150 models).

Target Segment	Target Type	True Architecture	Reference PDB	Mean Cosine ($\bar{c}$)	Job Outcome	Pre-Registered Rule Status
KLVFFA	Target	Antiparallel	3OW9	-0.935	5 of 5 jobs < 0.0	Correct (Antiparallel)

Target Segment	Target Type	True Architecture	Reference PDB	Mean Cosine ($\bar{c}$)	Job Outcome	Pre-Registered Rule Status
GAIIGL	Target	Antiparallel	3PZZ	+0.049	2 of 5 jobs < 0.0	Hedged (Failed criteria)
NKGAII	Control	Parallel	3Q2X	+1.000	5 of 5 jobs > +0.50	Correct (Parallel)
AIIGLM	Control	Parallel	2Y3J	+1.000	5 of 5 jobs > +0.50	Correct (Parallel)
VGSNKGAIIGL	Control	Parallel	5VOS	+0.984	5 of 5 jobs > +0.50	Correct (Parallel)
FAEDVGSNKGAIIGL	Control	Parallel	6OIZ	+0.798	5 of 5 jobs > +0.50	Correct (Parallel)

P6 produced the pre-registered "exactly one correct" outcome: KLVFFA correctly yielded 5 of 5 job means below 0.0 ($\bar{c}$ = -0.935), whereas GAIIGL failed the success criterion by producing only 2 of 5 job means below 0.0 ($\bar{c}$ = +0.049, hedged). All four parallel control segments met the success criterion with 5 of 5 job means exceeding +0.50. Under the pre-registered decision logic, this outcome narrows the study claim: AF3 does not fail uniformly on all Aβ antiparallel segments, but fails on some (MVGGVV, GAIIGL) while succeeding on another (KLVFFA). Post-hoc rationalizations for this divergence are precluded by the protocol.

Combining Phase 3 and P6 segments yields an updated 2 × 2 contingency matrix (Figure 3): non-Aβ systems were 3 of 3 correct for parallel and 2 of 2 for antiparallel; Aβ systems were 5 of 5 correct for parallel ([0.478, 1.000]) and 1 of 3 for antiparallel ([0.008, 0.906]). Secondary predictions confirmed that parallel controls retained fidelity (4/4), while interface confidence (`ipTM`) failed to separate correct from failing jobs (mean 0.351 vs. 0.220; Mann–Whitney U test, $p = 0.30$, $n = 27$ vs. $n = 3$).

Three structural observations are noted without post-hoc interpretation: KLVFFA includes three single-chain crystal structures (2Y29, 2Y2A, 9Q2V) whose in-sheet strand orientation cannot be evaluated; GAIIGL's failure was an architectural hedge where 24 of 25 diffusion samples contained an antiparallel bridge but 15 were classified as mixed; and FAEDVGSNKGAIIGL produced descriptor disagreements in 14 of 25 models while maintaining a parallel mean cosine of +0.798.

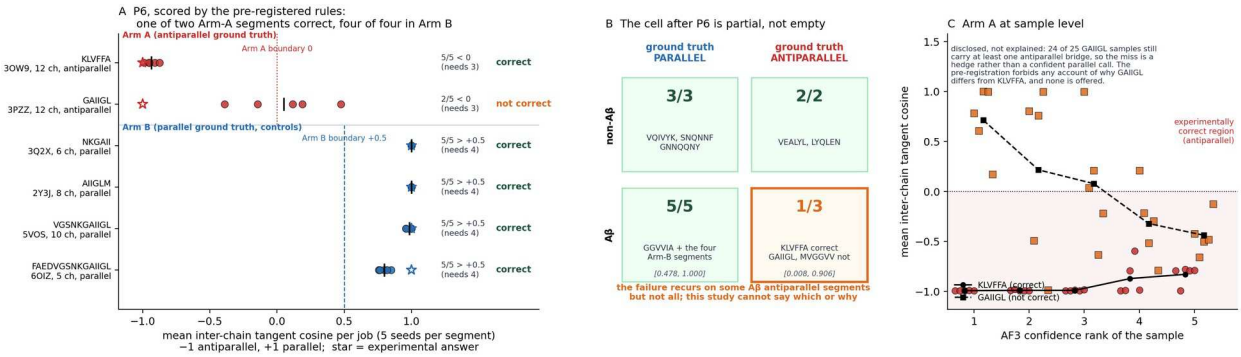

Figure 3. Pre-registered P6 benchmark and updated cross-system summary.

(A) Mean inter-chain tangent cosine per job (n = 5 seeds per segment) across six Aβ segments evaluated in P6. Open stars mark experimental reference values. Arm A evaluates antiparallel ground-truth targets (KLVFFA and GAIIGL; pre-registered success threshold at 0.0, red line); Arm B evaluates parallel ground-truth controls (NKGAII, AIIGLM, VGSNKGAIIGL, and FAEDVGSNKGAIIGL; pre-registered success threshold at +0.50, blue line). KLVFFA succeeded (5 of 5 jobs < 0.0) while GAIIGL hedged (2 of 5 jobs < 0.0), meeting the pre-registered outcome of exactly one correct target.

(B) Updated 2 × 2 contingency matrix summarizing segment-level architectural recovery across confirmatory Phase 3 and P6 benchmarks, categorized by sequence origin (non-Aβ vs. Aβ) and experimental ground truth (parallel vs. antiparallel), with exact 95% Clopper–Pearson confidence intervals.

(C) Individual diffusion sample cosines for Arm A targets (KLVFFA, red circles; GAIIGL, orange squares; n = 25 models per system) plotted across AF3 confidence ranks (Ranks 1 to 5). Black dashed line connects rank-averaged GAIIGL cosines, illustrating architectural decay toward negative values at lower confidence ranks despite failing the job-mean consensus threshold.

Source: Generated by p6_figure.py from p6_architecture.csv and p6_results.csv (30 jobs, 150 models). Figure created by the author.

3.6 Arm C: Polymorphic Sampling in MVGGVVIA (Exploratory)

Arm C evaluated the 8-residue Aβ segment MVGGVVIA across 5 seeds (25 models; 8 chains, templates disabled). This segment provides a polymorphic test case because both parallel (PDB 2Y3K) and antiparallel (PDB 2Y3L) crystal structures are experimentally deposited. By pre-registration, Arm C was designated strictly as an exploratory arm with no formal success criteria, and its outcomes are kept outside the confirmatory 2 × 2 classification matrix. Because both architectures represent experimentally validated conformations, neither parallel nor antiparallel predictions can be designated as correct or

incorrect; the primary observation lies in sample distribution across confidence ranks rather than binary accuracy.

AF3 generated both architectures across the ensemble, preferentially placing parallel packing at top ranks. All 5 top-ranked (Rank 1) models were cleanly parallel ($\bar{c}$ = +1.000). Mean directional cosines decreased monotonically across confidence ranks 1 through 4, with both Ranks 4 and 5 negative, against an overall system mean of +0.268. Across all 25 models, 7 reached a clean antiparallel consensus call, 6 of which occurred at Ranks 4 or 5. Because the pre-registration protocol specified no refutable prediction for this exploratory arm, no rank-order hypothesis tests were planned, and none are evaluated post-hoc.

3.7 Cross-Model Benchmark on Hexameric Aβ

To determine whether the parallel sheet preference observed in AlphaFold 3 is shared across other deep learning co-folding architectures, comparative predictions were evaluated using Boltz-1 (Wohlwend et al., 2024), Boltz-2 (Passaro et al., 2025), and Chai-1 (Chai Discovery, 2024). All three models were tested on the identical hexameric systems examined in AF3's template-free cohort: $A\beta_{1-40}$ wild-type, D23N (Iowa), and E22G (Arctic) across 6 chains, sampled over 5 pre-specified seeds with 5 diffusion samples per seed (15 jobs, 75 models per architecture; 225 models total). Scored using the identical `sheetarch` pipeline, local evaluation agreed with remote execution logs across all 225 models (Table 7). Zero indeterminate models were produced by any architecture across all 300 evaluations.

Table 7. Comparative architectural endpoints across deep learning models on hexameric $A\beta_{1-40}$ assemblies (WT, D23N, E22G; 15 jobs, 75 models per architecture, templates disabled).

Model Architecture	Jobs with Mean Cosine < 0.0	System Mean Cosine ($\bar{c}$)	Antiparallel Consensus Models	Models with ≥ 1 Antiparallel Bridge	Parallel Consensus Models	Mixed Models	Descriptor DISAGREEMENT
AlphaFold 3	0 of 15	+0.935	0 of 75	12 of 75	63 of 75	5 of 75	7 of 75
Boltz-1	0 of 15	+0.994	0 of 75	1 of 75	75 of 75	0 of 75	0 of 75
Chai-1	0 of 15	+0.671	0 of 75	25 of 75	40 of 75	7 of 75	28 of 75
Boltz-2	14 of 15	-0.322	19 of 75	68 of 75	5 of 75	31 of 75	20 of 75

The architectures displayed distinct topological profiles. Boltz-1 enforced parallel packing even more strictly than AF3, yielding a system mean cosine of +0.994 with all 75 models reaching parallel consensus. Chai-1 maintained a positive mean orientation ($\bar{c}$ = +0.671), but showed marked structural ambiguity: 28 of 75 models resulted in descriptor disagreements (`DISAGREEMENT`), including 16 of 25 wild-type samples. Conversely, Boltz-2 diverged from the other architectures, yielding negative job-mean cosines in 14 of 15 jobs ($\bar{c}$ = -0.322; -0.375 for WT, -0.486 for D23N, -0.103 for E22G) and antiparallel bridges in 68 of 75 models.

This divergence shows that an inductive parallel preference is not an inherent feature of all biomolecular diffusion models. However, Boltz-2's negative cosine does not reflect an accurate recovery of native fibrils: only 19 of 75 models reached a clean antiparallel consensus, while 51 were categorized as mixed (31) or descriptor disagreements (20). Outside of D23N, classical full-length wild-type and E22G fibrils are parallel in-register; the sole antiparallel wild-type deposition (PDB 7Y8Q) represents a membrane-associated non-fibrillar assembly rather than a cross-β fibril. A negative cosine in unconstrained hexamers cannot therefore be interpreted as native structural recovery.

4. Discussion and Limitations

4.1 Summary of Findings

This study tested whether AlphaFold 3 (AF3) accurately recovers the relative strand orientation (parallel versus antiparallel cross-β architecture) of amyloid assemblies across diverse structural systems. Across prospective and retrospective arms, the findings delineate a specific, bounded behavior:

1. AF3 does not possess a global inductive bias toward parallel β-sheets; this initial hypothesis was refuted in Phase 3 by positive antiparallel controls from insulin.
2. Prediction failure occurs on specific Aβ-derived antiparallel segments (MVGGVV and GAIIGL hedge) while another succeeds (KLVFFA), and full-length Aβ sampling never recovered the deposited antiparallel D23N polymorph.
3. Where partial antiparallel geometry is generated in full-length Aβ assemblies, it concentrates in lower-confidence, lower-ranked diffusion samples.

4.2 Interpretation of the Refutation and Boundary Narrowing

The empirical trajectory of this study refutes two broad generalizations regarding deep learning behavior on amyloid systems. First, topological mispredictions cannot be attributed to an intrinsic architectural incapacity: AF3 successfully generates cleanly antiparallel sheets for the steric-zipper segments VEALYL, LYQLEN, and KLVFFA. Second, the observed failures cannot be attributed to a blanket bias that forces all Aβ-derived sequences into parallel packing: all five parallel Aβ controls met the pre-registered success criterion, and the Aβ hexapeptide KLVFFA assembled into its native antiparallel topology across all five random seeds.

What remains is a localized breakdown across a subset of Aβ segments with antiparallel ground truth. Within this 1-of-3 cell (Figure 3), the wide exact confidence interval ([0.008, 0.906]) precludes asserting a formal population error rate.

The prospective design of this investigation constrained the theoretical claims that could be drawn from these data. Prior to executing the confirmatory P6 jobs, a retrospective narrative suggested that failure was governed strictly by the conjunction of an Aβ sequence context and an antiparallel ground truth. Under the frozen decision rules registered prior to job execution, the observation of "exactly one correct" outcome (KLVFFA correct, GAIIGL hedged) required the hypothesis to narrow: AF3 fails on some Aβ antiparallel segments, but not all. In accordance with the pre-registered protocol, we refrain from proposing post-hoc mechanistic rationalizations for why KLVFFA succeeded while GAIIGL and MVGGVV hedged. The value of pre-registration lies precisely in preventing the construction of retroactive explanations when an empirical boundary is encountered.

4.3 Evaluating Potential Explanations: Alignment Depth and Structural Census

Two exploratory hypotheses were evaluated to examine potential drivers of the observed Aβ failures:

Multiple Sequence Alignment (MSA) Depth. Because all seven Phase 3 steric-zipper segments possessed an effective alignment depth of 1, AF3's successful recovery of antiparallel insulin sheets occurred in the complete absence of evolutionary covariation signal. This rules out shallow alignments as the driver of architectural failure and confirms that MSA subsampling methods cannot be applied to these steric-zipper systems.

Training Set Structural Distribution. A census of experimental Aβ assemblies (≥ 4 chains) in the Protein Data Bank shows an asymmetric distribution: 93 of 97 entries are parallel, narrowing to 16 parallel versus 3 antiparallel entries among structures released on or before AF3's declared training cutoff (September 30, 2021). This pooled census mixes two structural classes, and the asymmetry is weaker in the class to which the failing targets belong: among Aβ fibrils the antiparallel share is 3 of 94 (96.8% parallel; exact 95% CI [0.007, 0.090]), whereas among Aβ segment crystals it is 7 of 52 (86.5% parallel; [0.056, 0.258]). MVGGVV, GAIIGL and KLVFFA are all segments, so the segment figure is the one relevant to the decisive cell of the contingency matrix, and it is the weaker of the two. A model trained on this corpus could plausibly learn a strong prior associating Aβ sequence motifs with parallel packing. However, this remains a candidate hypothesis rather than a proven mechanism: AF3's proprietary training data filtering is unverified, the pre-cutoff sample size is small (n = 19), the pooled figure is drawn from a population that mixes fibrils and segments, and a control census across other amyloid families is unavailable. Furthermore, this broad distribution cannot explain the divergent outcomes observed between short segments such as KLVFFA and GAIIGL.

4.4 Sampling, Confidence, and Practical Guidance

The retrospective evaluation of full-length $A\beta_{1-40}$ D23N illustrates a coverage failure on a polymorphic free-energy landscape. Experimentally, D23N forms both a stable parallel fibril (PDB 2MPZ) and a metastable antiparallel fibril (PDB 2LNQ; Qiang et al., 2012). Across 100 structural models spanning diverse template conditions, AF3 generated 96 parallel, 2 mixed, and 2 descriptor-disagreement models, never returning the antiparallel form even when supplied with an antiparallel monomeric template. AF3's output represents an incomplete sampling of accessible conformational space rather than an unphysical prediction.

Analyzing diffusion samples across confidence ranks reveals that partial antiparallel geometry rarely appears in top-ranked outputs. In the retrospective arms, antiparallel β-bridges were present in 24 of 160 jobs overall, but appeared in only 1 of 160 top-ranked models. For MVGGVV, the correct antiparallel architecture was generated but assigned to Rank 5 (the lowest confidence rank) under the calibrated cutoff (0.90); for D23N, the antiparallel polymorph was never generated in 20 jobs. Thus, stochastic sampling recovers lower-ranked native conformers for specific short segments, but did not rescue full-length fibril predictions.

Predicted confidence metrics do not provide a dependable check on topological validity. Across all 800 retrospective models, interface confidence (`ipTM`) never reached 0.80, and models displaying partial antiparallel character received systematically lower confidence scores (Δ = -0.059, $p = 6.3 \times 10^{-5}$). In P6, ipTM scores did not discriminate between correct and hedged predictions ($p = 0.30$).

Based on these structural and metric distributions, we offer three practical recommendations:

1. **Score all diffusion samples:** Analyses of multi-chain assemblies should evaluate all returned diffusion outputs rather than filtering exclusively on top-ranked models, as alternative quaternary topologies may concentrate in lower-confidence samples.
2. **Exercise caution when interpreting Aβ fibril variants:** Predictions of novel Aβ mutants or polymorphs should not be assumed to reflect physical quaternary architecture without orthogonal experimental validation.
3. **Do not treat confidence scores as architectural verification:** `ipTM` and `ranking_score` reflect internal model confidence estimates rather than checks on ground-truth architecture, and lower scores may accompany native metastable conformations.

4.5 Methodological Lessons in Descriptor Validation

Developing and benchmarking `sheetarch` yielded three methodological insights for automated structural evaluation:

- **Comprehensive validation across all callable classes:** Metric verification must include experimental ground truth for every categorical output state. Calibrating `sheetarch` against known parallel and antiparallel references caught an early bug that silently returned a parallel assignment when coordinate contacts were absent, establishing the principle of failing closed to an `indeterminate` call.
- **Segment-confined statistical evaluation:** Global rank-order statistics must not be pooled across discontinuous secondary structure elements. In PDB 2LNQ, calculating the register correlation across both segments of the hairpin fold produced an inverted correlation ($\rho = +0.429$), whereas segment-restricted evaluation correctly recovered antiparallel register ($\rho = -0.988$).
- **The job as the proper unit of analysis:** Treating individual diffusion samples from the same prediction job as independent observations introduces substantial pseudoreplication. For $A\beta_{1-40}$ D23N, model-level pooling shrinks the nominal upper bound below 0.036, whereas job-level aggregation yields a conservative bound below 0.168.

4.6 Limitations

1. **Small Confirmatory Sample Sizes:** The prospective Aβ antiparallel test cell contains three segments ($n = 3$), yielding wide confidence bounds ([0.008, 0.906]) that prevent asserting a population error rate; an exact binomial lower bound of 0.80 would require 17 consecutive successes.
2. **Retrospective Evaluation and Calibration Dependencies:** The 160 retrospective jobs were analyzed post-hoc. While the primary bridge criterion is parameter-free, the quantitative rank distribution for MVGGVV depends on the calibrated 0.90 upper mixed boundary, and the 15 reference structures reflect development calibration rather than held-out prospective validation.
3. **Template and Weight Access Constraints:** The public AlphaFold Server limits template inputs to single chains, preventing tests of whether multi-chain antiparallel templates direct fibril assembly. Furthermore, the web server does not expose internal checkpoint hashes; exact numerical reproduction requires local model weights.
4. **Unresolved KLVFFA Crystallographic Polymorphs:** PDB records for KLVFFA contain three single-chain crystal structures whose in-sheet strand orientation cannot be evaluated; an alternative parallel KLVFFA polymorph cannot be ruled out from the literature.
5. **Absence of Thermodynamic and Kinetic Modeling:** Chain counts were fixed by user inputs (2 to 12 chains across systems), and predictions represent static endpoints without modeling aggregation kinetics or free-energy landscapes.

6. **Unproven Mechanistic Basis:** The structural training-set skew remains an exploratory hypothesis; the specific sequence and geometric factors governing selective failure across Aβ segments remain unestablished.

4.7 Future Directions

Future work should evaluate additional antiparallel steric-zipper segments from non-Aβ amyloids to narrow confidence bounds, test multi-chain antiparallel packing templates using local AF3 weights, complete multi-family structural censuses across the PDB, and automate crystallographic symmetry expansion to classify single-chain steric-zipper depositions.

5. Conclusion

This study evaluated whether AlphaFold 3 accurately determines quaternary amyloid β-sheet architecture or defaults to an inductive parallel bias. Using `sheetarch`, a fail-closed geometric descriptor suite validated across 15 experimental structures, we subjected model predictions to two pre-registered confirmatory benchmarks alongside large-scale retrospective evaluations.

The findings refute a global parallel prediction bias: AF3 recovers antiparallel β-sheets for non-Aβ controls and parallel β-sheets across all tested systems. However, its performance narrows on Aβ-derived sequences, where it fails on specific antiparallel segments (MVGGVV and GAIIGL) while succeeding on another (KLVFFA), and consistently fails to recover the deposited antiparallel polymorph of full-length $A\beta_{1-40}$ D23N. In full-length assemblies, partial antiparallel geometry is concentrated in lower-confidence diffusion samples, meaning that evaluating only top-ranked models conceals alternative topological states.

Practitioners should not rely on AF3's confidence scores or top-ranked models to infer quaternary sheet architecture in polymorphic amyloid systems without experimental validation. Because every confirmatory hypothesis and decision boundary was committed in advance, the empirical narrowing of these findings illustrates the value of pre-registered protocols when evaluating deep learning frameworks on complex macromolecular assemblies.

References

Abramson, J., et al. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016), 493–500.

Chai Discovery. (2024). Chai-1: Decoding the molecular interactions of life. Technical Report.

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4), 404–413.

Colletier, J.-P., et al. (2011). Molecular basis for amyloid-beta polymorphism. *Proceedings of the National Academy of Sciences*, 108(41), 16938–16943.

del Álamo, D., et al. (2022). Sampling alternative conformational states of transporters and receptors with AlphaFold2. *eLife*, 11, e75751.

Jumper, J., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589.

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12), 2577–2637.

Landau, M., et al. (2011). Towards a pharmacophore for amyloid. *PLOS Biology*, 9(6), e1001080.

Mirdita, M., et al. (2022). ColabFold: Making protein folding accessible to all. *Nature Methods*, 19(6), 679–682.

Passaro, S., et al. (2025). Boltz-2: Towards accurate and general biomolecular interaction modeling. *bioRxiv*, preprint.

Purushothaman, G. (2026). Replication-based evaluation of AlphaFold3 for predicting mutation effects in two Alzheimer's-disease protein systems: A benchmark against an experimental amyloid-beta fibril structure. In Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Dallas, TX. In press.

Qiang, W., et al. (2012). Antiparallel beta-sheet architecture in Iowa-mutant beta-amyloid fibrils. Proceedings of the National Academy of Sciences, 109(12), 4443–4448.

Sawaya, M. R., et al. (2007). Atomic structures of amyloid cross-beta spines reveal varied steric zippers. *Nature*, 447(7143), 453–457.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6), 657–680.

Tycko, R. (2015). Amyloid polymorphism: Structural basis and neurobiological relevance. *Neuron*, 86(3), 632–645.

Wayment-Steele, H. K., et al. (2024). Predicting multiple conformations via sequence clustering and AlphaFold2. Nature, 625(7996), 832–839.

Wohlwend, J., et al. (2024). Boltz-1: Democratizing biomolecular interaction modeling. *bioRxiv*, preprint.