When to Trust AlphaFold: A Trajectory-Aware Evaluation of ColabFold on CASP14 Targets

Riyaz Ahuja

Carnegie Mellon University
riyaza@andrew.cmu.edu

Abstract

Predicted protein structures are now routine inputs to biological reasoning, but a coordinate file alone does not establish when the prediction should be trusted. This paper evaluates that reliability problem using ColabFold 1.6.1 on 30 curated CASP14-derived targets with experimentally determined references. We frame monomer structure prediction as risk minimization over chemically feasible coordinate sets and interpret the AlphaFold2 model family, as run through ColabFold, as an amortized generator that returns coordinates together with confidence estimates, ranked alternatives, and recycle snapshots. The benchmark reproduces strong AlphaFold2-style accuracy: 29/30 final rank-1 predictions reach topology-level success, with median TM-score 0.862. The reliability analysis is more nuanced. Mean pLDDT is the strongest aggregate signal of global accuracy in this benchmark, recycling produces a statistically reliable average improvement but is not monotone for every target, rank-1 is usually very close to the best of five models, and MSA depth is most informative in extreme cases rather than as a significant aggregate correlation. The main conclusion is therefore interpretive: AlphaFold2-style predictions are best treated as confidence-scored model outputs with alternative ranked structures and refinement behavior, not as single static experimental structures.

1 Introduction

A protein sequence is a compact genetic description; biological function is expressed in three dimensions. Local chemistry, secondary structure, residue packing, active-site geometry, binding interfaces, and allosteric couplings all depend on where atoms sit in space. The central biological problem is that sequences are easy to read at scale, while the functionally relevant three-dimensional structures are expensive to measure. X-ray crystallography, NMR spectroscopy, and cryo-EM remain the empirical standards, and the Protein Data Bank provides the reference archive for solved structures (Berman et al. 2000). Yet experimental coverage is sparse compared with the scale of modern sequence databases, so computational structure prediction has become a core problem in computational biology.

The biological statement of the problem is subtle. A protein sequence does not determine a single rigid object in isolation. It determines, together with solvent, pH, ligands, binding partners, oligomeric state, and cellular context, a distribution over conformations. Experimental structures are snapshots or summaries of that distribution under particular conditions. Nevertheless, for many folded monomeric domains there is a dominant native state, and predicting a representative structure is extraordinarily useful for interpreting function, designing experiments, identifying binding sites, and reasoning about mutation effects. The goal of a structure predictor is therefore not merely to produce a visually plausible ribbon diagram, but to infer a chemically feasible geometry that is close enough to the relevant biological state to support downstream reasoning.

AlphaFold2 changed the practical status of this problem by producing near-experimental accuracy for many CASP14 targets (Jumper et al. 2021; Kryshtafovych et al. 2021). ColabFold made this capability easier to run locally and at scale by packaging AlphaFold2-style models with fast MMseqs2-based multiple-sequence-alignment generation (Mirdita et al. 2022; Steinegger and Söding 2017). The AlphaFold Protein Structure Database then made predicted structures and confidence estimates available at proteome scale (Varadi et al. 2022). As a result, structure prediction is no longer only an algorithmic challenge. It is also a reliability challenge. A user must know when a predicted structure is accurate enough for the biological claim being made.

Throughout this paper, AlphaFold2 refers to the neural architecture, learned model family, and confidence machinery introduced by Jumper et al. (2021). ColabFold refers to the implementation used in this project: the practical workflow that generates MSAs, runs AlphaFold2-PTM model parameters, writes PDB and confidence outputs, and ranks the resulting predictions.

This project evaluates that reliability question on a controlled benchmark. We ran ColabFold on 30 CASP14-derived targets with processed experimental references, disabled templates, saved all five model outputs, and evaluated recycle snapshots as well as final structures. The standard replication question is whether ColabFold recovers the correct folds. The more important question is when its outputs should be trusted. The analyses therefore connect AlphaFold2's model mechanisms to observable ColabFold diagnostics: MSA depth is tied to evolutionary evidence, recycling is tied to iterative refinement, pLDDT and PAE are

tied to confidence estimation, and model ranking is tied to the choice of a final prediction.

The paper is organized as follows. Section 2 reviews related methods for protein structure prediction and explains what changed with AlphaFold2. Section 3 formalizes the structure prediction task and gives a mathematical description of the AlphaFold2 model and the ColabFold inference workflow as a conditional structure generator. Section 4 describes the dataset, preprocessing, inference, and evaluation algorithms. Section 5 presents aggregate results, with interpretation interleaved into each analysis. Section 6 gives detailed case studies of a success, a failure, and a high-confidence mismatch. Section 7 concludes, and the appendix gives reproducibility details.

2 Related Work

Comparative modeling. Classical protein structure prediction first succeeded most reliably when a homologous structure was already known. Comparative or homology modeling identifies a template structure, aligns the query sequence to that template, copies or restrains conserved backbone geometry, and models insertions, deletions, side chains, and loops (Sali and Blundell 1993). This approach is powerful because evolutionary relatedness often preserves fold topology. Its main weakness is the same dependency that makes it powerful: if no good template exists, or if the sequence-template alignment is wrong, the predicted model can inherit incorrect geometry. Template-based modeling also tends to struggle with novel folds, flexible regions, and conformational changes.

Threading and fold recognition. Threading methods extended template reasoning by asking whether a sequence can fit any fold in a structural library, even when obvious sequence homology is weak. A sequence is scored against known three-dimensional folds using environmental preferences, residue-residue compatibility, and structural profiles (Bowie, Luthy, and Eisenberg 1991). Threading is useful because the number of protein folds is smaller than the number of sequences, so a remote fold match can be informative. But threading remains bounded by the fold library and by the scoring function's ability to distinguish a true remote homolog from a misleading structural analogy.

Fragment assembly and energy search. Fragment-based methods such as Rosetta model structure prediction as a search over conformations guided by fragment libraries and physically inspired energy functions (Rohl et al. 2004). The algorithm samples local fragments, assembles them into global structures, and optimizes a score that approximates favorable packing, hydrogen bonding, torsions, solvation, and sterics. This framed structure prediction explicitly as an optimization problem over a rough energy landscape. The difficulty is computational: the conformational space grows rapidly with sequence length, and hand-designed energy functions are imperfect surrogates for biological stability.

Hybrid and meta-server approaches. Methods such as I-TASSER combine template identification, threading, fragment assembly, structure reassembly, and iterative refinement (Yang et al. 2015). These systems made practical gains by integrating multiple weak signals rather than relying on a single model of folding. They also foreshadowed a central lesson of modern structure prediction: useful predictions come from combining sequence evidence, structural priors, and geometric refinement. However, these systems still relied heavily on explicit template libraries, hand-engineered scoring functions, and complex sampling procedures.

Coevolution and deep learning before AlphaFold2. The growth of sequence databases made coevolutionary analysis a major source of structural information. If two residues mutate in a correlated way across homologous proteins, that coupling can indicate spatial proximity. Direct coupling methods and sparse inverse-covariance approaches showed that evolutionary variation can be converted into contact constraints (Marks et al. 2011; Jones et al. 2012). Later deep learning methods predicted distances, contacts, orientations, and potentials from MSAs and sequence features (Senior et al. 2020; Yang et al. 2020). These methods were a major step because they learned structural constraints instead of relying only on manually specified potentials.

End-to-end learned structure predictors. AlphaFold2 differs from earlier methods not because it ignores the historical ingredients, but because it integrates them in a learned end-to-end system. The Evoformer processes both MSA information and residue-pair representations; the structure module maps learned representations into coordinates; recycling feeds outputs back into the network for refinement; and confidence heads predict pLDDT and PAE (Jumper et al. 2021). RoseTTAFold reached similar territory with a three-track network that exchanges information among sequence, residue-pair, and coordinate representations (Baek et al. 2021). More recently, protein language-model approaches such as OmegaFold and ESMFold showed that large sequence models can support fast MSA-free structure prediction, although MSA-based systems remain especially strong when homologous sequence information is available (Wu et al. 2022; Lin et al. 2023). RoseTTAFold2 further blended ideas from RoseTTAFold and AlphaFold2, including frame-aligned error, recycling, and large-scale distillation (Baek et al. 2023). AlphaFold3 later moved the field beyond monomeric protein structure prediction toward diffusion-based prediction of broader biomolecular complexes, including proteins, nucleic acids, ligands, ions, and modified residues (Abramson et al. 2024). Those systems are outside the scope of this benchmark, but they reinforce the same reliability question: modern structure predictors output plausible geometries with confidence estimates, not experimental truth.

ColabFold and prediction at scale. ColabFold keeps the core AlphaFold2 modeling idea while making inference practical through fast MMseqs2-based MSA generation and a simplified user workflow (Mirdita et al. 2022; Steinegger and Söding 2017). Public resources such as AlphaFold DB further changed the use case: predicted structures are now routine inputs to biological reasoning, not rare outputs from

specialized modeling groups (Varadi et al. 2022). Compared with classical methods, AlphaFold2-style prediction is less a hand-coded search algorithm and more an amortized inference system: training learns a mapping from sequence-derived features to structures, and inference applies that mapping to a new sequence.

This comparison motivates the present work. If a prediction is produced by a confidence-scored generator rather than a transparent physical simulation, then trust must be evaluated through diagnostics. MSA depth, predicted local confidence, predicted aligned error, model agreement, ranking behavior, and recycle dynamics are not side details. They are the observable traces of the information sources and refinement mechanisms that created the structure.

3 Structure Prediction and AlphaFold2

3.1 Biological problem statement

Let $S = (s_1, \ldots, s_L)$ be a protein sequence over the amino-acid alphabet $\mathcal{A}$. Biologically, S specifies a polymer whose possible conformations depend on covalent chemistry, sterics, solvent, temperature, binding partners, ligands, post-translational modifications, oligomeric state, and cellular context. Let E denote that environment and let X be a coordinate set for the atoms of the protein. The most faithful target is not a single static object but a conditional distribution over conformations,

$$p^\star(X \mid S, E).$$

A thermodynamic idealization writes this distribution as

$$p^\star(X \mid S, E) = \frac{1}{Z(S,E)} \exp[-\beta G(X; S, E)],$$

where G is the free energy, $\beta = (k_B T)^{-1}$, and Z is the partition function. Structure prediction would be solved if one could evaluate $G(X; S, E)$ accurately and search over all feasible conformations. In practice, both the energy function and the conformational search problem are intractable.

Benchmarks simplify this biological target. They assume that a target domain has an experimentally observed reference structure Y that is a useful representative of the relevant native state. The computational task then becomes: given S and allowed auxiliary information, output a chemically plausible coordinate set $\hat{X}$ that is close to Y up to rigid motion and unresolved residues. This simplification is necessary for evaluation, but it matters for interpretation. A prediction can disagree with one reference conformer because it is wrong, or because the protein is flexible, multi-state, bound differently, or measured under conditions that differ from the prediction context.

3.2 Mathematical formulation

Coordinate space and feasibility. For a sequence S, define a coordinate space

$$\mathcal{X}(S) \subset \mathbb{R}^{N(S)\times 3},$$

where $N(S)$ is the number of modeled atoms. The feasible set $\mathcal{X}(S)$ is constrained by chain connectivity, bond lengths, bond angles, chirality, excluded volume, residue chemistry, and, when side chains are modeled, rotameric constraints. Since absolute position and orientation are arbitrary, structures are compared in the quotient space $\mathcal{X}(S)/\mathrm{SE}(3)$, for the special Euclidean group $\mathrm{SE}(3)$ of rigid motions.

Structure prediction as optimization. Let $\ell(X, Y)$ be a structural loss invariant to global rotation and translation. The ideal point prediction minimizes expected risk under the native conformational distribution:

$$X^\star(S, E) = \operatorname*{arg\,min}_{X \in \mathcal{X}(S)/\mathrm{SE}(3)} \mathbb{E}_{Y \sim p^\star(\cdot \mid S,E)}[\ell(X, Y)].$$

Different losses correspond to different biological goals. RMSD emphasizes coordinate agreement over aligned atoms. $1 - \mathrm{TM}$ emphasizes global fold topology. Local distance losses emphasize residue-neighborhood reliability. A structure useful for fold annotation may be insufficient for active-site chemistry or domain-placement claims.

In a benchmark with one experimental reference Y_i per sequence, the empirical structure prediction problem is

$$\min_f \frac{1}{n} \sum_{i=1}^{n} \ell\big(f(S_i, A_i, T_i), Y_i\big),$$

where A_i is a multiple sequence alignment, T_i is optional template information, and f is a prediction algorithm. This project evaluates this empirical objective with several choices of ℓ and with additional confidence-calibration diagnostics. AlphaFold2 is not explicitly optimizing that benchmark objective at inference time; it has learned an amortized mapping whose outputs are expected to have low structural loss on new sequences.

Evaluation metrics and homology quantities. The primary global metric is reference-normalized TM-score from US-align (Zhang et al. 2022). TM-score measures structural similarity while reducing the length bias of RMSD (Zhang and Skolnick 2004). After structural alignment,

$$\mathrm{TM} = \max_{R,t} \frac{1}{L_{\mathrm{ref}}} \sum_i \frac{1}{1 + (\|Rx_i + t - y_i\|/d_0(L_{\mathrm{ref}}))^2},$$

where x_i and y_i are aligned residue coordinates, R, t are a rigid superposition, and d_0 is a length-dependent scale. TM-score near 1 indicates close structural agreement; TM-score ≥ 0.5 is commonly treated as topology-level success.

RMSD is

$$\mathrm{RMSD} = \min_{R,t} \sqrt{\frac{1}{N} \sum_{i=1}^{N} \|Rx_i + t - y_i\|^2}.$$

It has the intuitive unit of Å, but is sensitive to outliers, domain motion, and alignment coverage. The local C_α error used in this paper is

$$e_i = \|Rx_i + t - y_i\|,$$

computed after global superposition for each aligned residue.

For homology support, raw MSA depth is the number of aligned sequences. The approximate effective sequence count is

$$Neff = \sum_{a=1}^{M} \frac{1}{|\{b : \mathrm{id}(a,b) \geq 0.8\}|},$$

which downweights redundant homologs. Neff is therefore a diversity-weighted measure of evolutionary information.

3.3 AlphaFold2 and ColabFold as an ML system

The architectural description in this section follows the AlphaFold2 model and methods described by Jumper et al. (2021). ColabFold-specific inference details, especially the use of MMseqs2 for fast MSA generation, follow Mirdita et al. (2022) and Steinegger and Söding (2017).

Inputs and featurization. For a query sequence S, ColabFold constructs an MSA

$$A = q_D(S) \in (\mathcal{A} \cup \{\mathrm{gap}\})^{M \times L}$$

using MMseqs2 over sequence database D. The first row is the query sequence and the remaining rows are homologs. Optional template features are denoted T; in this project $T = \varnothing$. AlphaFold2 embeds these inputs into two main tensors:

$$M^0 \in \mathbb{R}^{M \times L \times d_m}, \qquad Z^0 \in \mathbb{R}^{L \times L \times d_z}.$$

M^0 is the MSA representation, storing information by homolog and residue position. Z^0 is the pair representation, storing information about ordered residue pairs. It includes sequence separation, residue identity information, template-derived geometry when templates are enabled, and information passed from the MSA through later Evoformer blocks.

Evoformer representations. The Evoformer updates the MSA and pair tensors through K blocks:

$$(M^k, Z^k) = E_k(M^{k-1}, Z^{k-1}), \qquad k = 1, \ldots, K.$$

Each block combines several operations. Row-wise MSA attention lets each sequence position attend across residues, using the pair representation as a geometric bias. Column-wise MSA attention lets homologous residues at the same position exchange information across sequences. An outer-product-mean operation transfers coevolutionary signal from M^k into pair features Z^k. Triangle multiplicative updates and triangle attention then update pair features through residue triples (i, j, k), encouraging geometric consistency. Schematically,

$$M^k \leftarrow M^{k-1} + \mathrm{MSAAttn}(M^{k-1}; Z^{k-1}),$$
$$Z^k \leftarrow Z^{k-1} + \mathrm{OPM}(M^k) + \mathrm{Triangle}(Z^{k-1}).$$

This is the core learned reasoning stage: evolutionary patterns in the MSA are converted into residue-pair constraints that can support a global fold.

Structure module and recycling. The structure module reads the final single and pair representations. Let s_i be the single representation for residue i, derived from the query row of the MSA representation. The structure module uses invariant geometric operations to predict residue frames $F_i \in \mathrm{SE}(3)$ and torsion angles χ_i:

$$(F_i, \chi_i)_{i=1}^{L} = \mathcal{S}_\theta\big((s_i)_{i=1}^{L}, Z^K\big).$$

All-atom coordinates are then constructed by applying residue-specific geometry to these frames and torsions:

$$X = g\big((F_i, \chi_i)_{i=1}^{L}, S\big).$$

The key operation inside the structure module is invariant point attention (IPA). Ordinary attention passes scalar features between residues, but coordinate generation also requires geometric consistency: rotating or translating the input coordinate frame should rotate or translate the output structure in the same way, without changing the internal prediction. IPA achieves this by maintaining a local rigid frame for each residue and exchanging both scalar information and learned points expressed in those frames. Distances between points are invariant to global rigid motion, while the frames themselves carry orientation information needed to place the backbone. This frame-based construction is why AlphaFold2 can move from pairwise constraints to a coherent 3D structure without treating coordinates as arbitrary Euclidean vectors.

Recycling repeats the network with information from the previous prediction. At recycle r, the model has coordinates $X^{(r)}$ and final representations from the previous pass. These are embedded and added back into the next pass:

$$(M^{0,(r+1)}, Z^{0,(r+1)}) = (M^0, Z^0) + \mathrm{RecycleEmbed}\big(X^{(r)}, M^{K,(r)}, Z^{K,(r)}\big).$$

The next recycle produces $X^{(r+1)}$. Recycling is therefore a learned refinement iteration, not a physical folding trajectory and not a guaranteed descent step on TM-score or free energy.

Confidence heads: pLDDT and PAE. AlphaFold2 computes confidence values with prediction heads attached to its learned representations. The pLDDT head maps each residue's final single representation to a categorical distribution over lDDT bins. lDDT is a local distance-difference measure of whether a residue's neighboring interatomic distances are preserved (Mariani et al. 2013); AlphaFold2 predicts it without seeing the reference structure:

$$\pi_i = \mathrm{softmax}(W_{\mathrm{lddt}} s_i + b_{\mathrm{lddt}}).$$

If c_b is the center of lDDT bin b, the reported confidence is the expected bin value scaled to 0–100:

$$pLDDT_i = 100 \sum_b c_b \, \pi_{i,b}.$$

This is why pLDDT should be interpreted locally: it is computed per residue from a head trained to predict local

distance-difference accuracy (Mariani et al. 2013; Jumper et al. 2021).

The PAE head maps each final pair representation Z_{ij}^K to a categorical distribution over aligned-error bins:

$$\eta_{ij} = \mathrm{softmax}(W_{\mathrm{pae}} Z_{ij}^K + b_{\mathrm{pae}}).$$

If a_b is the center of aligned-error bin b, then

$$PAE_{ij} = \sum_b a_b \, \eta_{ij,b}.$$

Because PAE is pairwise, it can expose uncertainty in relative residue or domain placement even when local pLDDT values are high. Conceptually, PAE_{ij} estimates the position error of residue j when the predicted and true structures are aligned on residue i, using only the model's internal representations.

Training objective. AlphaFold2 is trained by supervised learning on known structures, with auxiliary self-supervised and geometric losses. A simplified objective is

$$\min_\theta \; \mathbb{E}_{(S,A,T,Y)} \big[\lambda_F \mathcal{L}_{\mathrm{FAPE}} + \lambda_\chi \mathcal{L}_\chi + \lambda_D \mathcal{L}_{\mathrm{dist}} + \lambda_M \mathcal{L}_{\mathrm{MSA}} + \lambda_C \mathcal{L}_{\mathrm{conf}} + \lambda_V \mathcal{L}_{\mathrm{viol}} \big].$$

The frame-aligned point error (FAPE) compares predicted and true atom positions in local residue frames:

$$\mathcal{L}_{\mathrm{FAPE}} \approx \frac{1}{L^2} \sum_{i,j} \min \left(\| F_i^{-1} x_j - \hat{F}_i^{-1} \hat{x}_j \|, d_{\mathrm{clamp}} \right),$$

where hats denote predicted frames and points. The distogram loss supervises pairwise distance distributions, the torsion loss supervises side-chain and backbone angles, the masked-MSA loss trains sequence-representation learning, the violation loss penalizes chemically implausible structures, and the confidence losses train the pLDDT and PAE/aligned-error heads. The trained network then amortizes structure optimization: it learns parameters that make one forward pass plus recycling approximate a low-loss solution for new sequences.

The intuition behind FAPE is that the model is penalized for placing atoms incorrectly relative to many residue-centered coordinate frames, not just after one global superposition. This makes the loss sensitive to local geometry, relative orientation, and long-range arrangement while remaining well behaved under global rigid motion.

ColabFold inference and ranking. For inference, ColabFold evaluates several AlphaFold2-PTM model parameterizations. Let ζ_m denote the model choice and implementation randomness for generated model m. A compact inference map is

$$(X_m, C_m) = G_\theta(S, q_D(S), T; \zeta_m),$$

where C_m contains pLDDT, PAE, and ranking metadata. The output set

$$\{(X_m, C_m)\}_{m=1}^5$$

is an ensemble-like sample from the model system. In this monomer benchmark, the PTM variant is used because it exposes the predicted aligned-error and predicted-TM machinery in addition to local pLDDT. PTM stands for predicted TM-score: the model estimates a TM-score-like global confidence quantity from its pair/aligned-error predictions. This does not change the task into protein-complex prediction; it gives the monomer run additional global and pairwise confidence diagnostics that are useful for trustworthiness analysis. ColabFold ranks the resulting monomer predictions primarily by confidence, which in this setup is operationally close to mean pLDDT:

$$\hat{m} = \arg\max_m \frac{1}{L} \sum_{i=1}^{L} pLDDT_{m,i}.$$

The rank-1 model is therefore the most confident generated model, not necessarily the model with maximal reference TM-score. This distinction is a central reason for evaluating rank-1 versus best-of-five behavior.

Mechanisms relevant to trust. The mathematical model determines the diagnostics used in this paper. MSA depth and Neff probe the evolutionary evidence entering M^0. TM-score and RMSD measure reference-based structural loss. Recycling trajectories evaluate the refinement iteration. Rank-1 versus best-of-five tests whether the confidence ranking matches reference accuracy. pLDDT is evaluated against local C_α error because it is a local confidence head. PAE is used as a warning signal for relative placement because it is a pairwise aligned-error estimate. The benchmark is therefore not just a final-output accuracy test; it asks whether the model's inputs, internal mechanisms, and confidence heads explain when predictions succeed or fail.

4 Dataset and Methodology

4.1 Benchmark

The benchmark contains 30 CASP14-derived targets with experimentally determined references from RCSB PDB. The set includes 4 TBM-easy targets, 7 TBM-hard targets, 4 mixed FM/TBM targets, 13 FM targets, and 2 multi-domain controls. All aggregate results use all 30 targets. Lengths range from 72 to 621 amino acids. Most targets are below 300 residues; a few longer and multi-domain examples serve as stress cases.

Category	n	Length range	Median TM-score
TBM-easy	4	193–621	0.954
TBM-hard	7	119–178	0.869
FM/TBM	4	72–102	0.892
FM	13	92–404	0.825
MultiDom	2	190–273	0.814
All	30	72–621	0.862

Table 1: Benchmark composition. The median TM-score column is included here to show that the target categories differ in difficulty, but all reported analyses use the full 30-target set.

4.2 Reference preprocessing

Reference preprocessing is necessary because CASP domains, PDB chains, and resolved experimental residues do

not always match exactly. Comparing a prediction to unresolved residues would produce artificial errors, while including residues outside the target domain would mix the intended task with extra structure. Algorithm 1 gives the full preprocessing and inference procedure. Conceptually, the reference step maps each CASP target domain onto the resolved residues in the corresponding PDB chain, removes unresolved or out-of-domain residues, and writes a matched FASTA/PDB pair. This produces an evaluation pair (S_i, Y_i) in which the input sequence and reference structure cover the same observed domain. Most targets are fully or nearly fully resolved. The largest missing-residue case is T1036s1-D1, where 583 residues are resolved out of 621 expected residues.

4.3 Inference and trajectory extraction

Inference used ColabFold 1.6.1 with AlphaFold2-PTM, five models per target, three recycle iterations, no templates, and no Amber relaxation. The PTM model family was chosen because it exports PAE and predicted-TM-style global confidence information in addition to local pLDDT, which is useful for a trustworthiness analysis. Templates were disabled to avoid using released structures as direct modeling information. Amber relaxation was skipped because the study focuses on fold topology, backbone geometry, confidence, and recycling rather than detailed steric minimization.

For each target, ColabFold first generated an MSA using MMseqs2, then ran the five model parameterizations. The pipeline saved final unrelaxed PDBs, confidence files, ranking metadata, PAE arrays, and recycle PDB snapshots. With 30 targets, 5 models, and 4 recycle endpoints (0 through 3), the trajectory component contains 600 recycle structures. Including final structures, the evaluation contains 750 structure-confidence records. Algorithm 1 summarizes these steps.

4.4 Evaluation algorithms

The evaluation pipeline separates global structural accuracy, confidence calibration, model ranking, and recycle dynamics. For every evaluated structure, US-align is run against the processed reference. The parser records reference-normalized TM-score, RMSD, and aligned length. Confidence parsers read per-residue pLDDT values and PAE arrays from ColabFold outputs. The local calibration step extracts matched C_α coordinates, applies a Kabsch superposition, computes residue-level C_α errors, and joins those errors to pLDDT values. Algorithm 2 summarizes the evaluation procedure.

The local-error table contains 129,350 residue-level rows. It intentionally pools targets, all five models, recycle snapshots, and final snapshots. The absolute local-error medians from this pooled table are therefore not final rank-1 error estimates; they are diagnostic values for testing whether higher pLDDT corresponds to lower aligned residue error under broad sampling.

Algorithm 1 Reference preparation and ColabFold inference

Require: Target table $\mathcal{B}$, sequence database D, PDB references
Ensure: Processed references, MSAs, final models, recycle snapshots
  for target i in $\mathcal{B}$ **do**
    Read target sequence S_i, domain boundaries, PDB id, and chain id.
    Download and parse the reference mmCIF structure.
    Extract resolved residues for the requested chain.
    Align S_i to the resolved chain sequence and map domain positions.
    Drop residues outside the domain or missing atomic coordinates.
    Write processed FASTA $\tilde{S}_i$ and processed reference Y_i.
    Construct MSA $A_i = q_D(\tilde{S}_i)$ using ColabFold/MMseqs2.
    Compute raw MSA depth and approximate Neff from A_i.
    for model $m = 1, \ldots, 5$ **do**
      Run AlphaFold2-PTM with templates disabled and recycle limit $R = 3$.
      Save $X_{i,m,r}$ for recycle endpoints $r = 0, 1, 2, 3$.
      Save final $X_{i,m,R}$, pLDDT, PAE, and ranking metadata.
    end for
  end for

Algorithm 2 Accuracy, confidence, ranking, and trajectory evaluation

Require: Processed references Y_i, predictions $X_{i,m,r}$, confidences $C_{i,m,r}$
Ensure: Aggregate accuracy, confidence, ranking, and local-error tables
  for each target/model/snapshot tuple (i, m, r) **do**
    Align $X_{i,m,r}$ to Y_i with US-align.
    Parse TM-score, RMSD, aligned length, pLDDT, and PAE.
    Store a row in the trajectory table.
  end for
  for each target i **do**
    Compare ColabFold rank to reference-based final-model TM-score rank.
    Compute rank-1 gap: $\max_m \mathrm{TM}_{i,m} - \mathrm{TM}_{i,\hat{m}}$.
  end for
  for aligned prediction-reference pair **do**
    Extract C_α pairs and estimate (R, t) by Kabsch superposition.
    Compute local error $e_j = \|Rx_j + t - y_j\|$ for aligned residue j.
    Join e_j to residue-level pLDDT_j and metadata.
  end for
  Aggregate by target, category, recycle step, confidence bin, and failure flag.

4.5 Reproducibility

The accompanying GitHub repository[1] contains the scripted workflow, processed references, saved ColabFold outputs, evaluation tables, and figure-generation code. The run used ColabFold 1.6.1 with AlphaFold2-PTM, five models per target, three recycles, no templates, and no Amber relaxation on two RTX 6000 Ada GPUs. The main evaluation convention is that references are processed before comparison, so unresolved residues are not counted as prediction errors. For US-align, the reported TM-score is normalized by the processed reference structure.

The main caveat is that ColabFold's public MMseqs2 service and underlying sequence databases can change over

[1] https://github.com/riyazahuja/TrustAlphaFold

time, which may alter generated MSAs and therefore predictions. The saved structures, confidence files, and evaluation tables are the stable artifacts for reproducing the analyses reported here.

5 Results

5.1 Final prediction accuracy

The first question is whether the ColabFold run reproduces the expected AlphaFold2-style baseline before asking more detailed trustworthiness questions. It does. The final rank-1 prediction reaches topology-level accuracy (TM-score $\geq$ 0.5) for 29 of the 30 targets, with median TM-score 0.862 and median RMSD 2.08 Å. This establishes that the later analyses are not mostly explaining a failed benchmark; they are probing why a generally strong predictor is more reliable in some regimes than others. The endpoints anchor the range: T1025-D1 is nearly exact (TM-score 0.989, RMSD 0.70 Å), whereas T1043-D1 is the only topology-level failure (TM-score 0.313) and aligns only 82 of 148 residues by US-align.

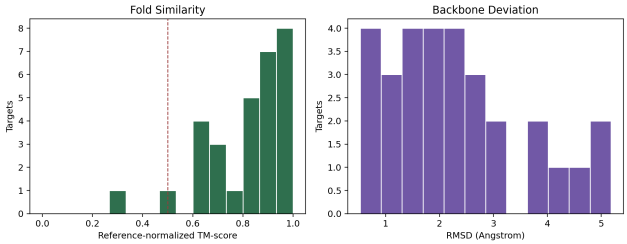


Figure 1: Final rank-1 benchmark accuracy. ColabFold recovers topology-level accuracy for 29/30 targets, with one clear low-TM-score failure.

The category trend is consistent with the biological difficulty gradient. The TBM-easy targets have median TM-score 0.954, TBM-hard targets have median TM-score 0.869, FM/TBM targets have median TM-score 0.892, FM targets have median TM-score 0.825, and the two multi-domain controls have median TM-score 0.814. This should not be overinterpreted as a perfectly monotone ranking because the sample sizes are small and the benchmark intentionally contains stress cases. Nevertheless, the main pattern is coherent: template-based and mixed-regime targets are usually high quality, while FM targets are more variable and contain the major failure. Full per-target values are reported in Appendix A.

5.2 MSA depth and difficulty

Because AlphaFold2 uses evolutionary information, MSA support should explain some of the accuracy variance. In this benchmark, that expectation is best treated as a mechanistic interpretation of the extreme cases rather than a statistically confirmed aggregate relationship. The scatter in Figure 3 has a positive but non-significant Spearman correlation ($\rho = 0.29$, $p = 0.12$), so the data do not justify claiming that MSA depth alone explains accuracy at $n = 30$. The clearest failure, T1043-D1, has the shallowest MSA: 6 sequences

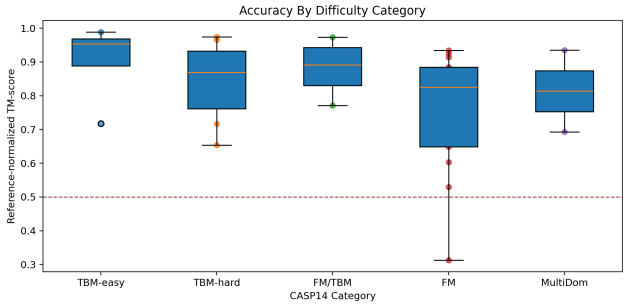


Figure 2: Accuracy by CASP category. FM targets show the widest spread and contain the only topology-level failure.

and approximate Neff 2.0. It also has low mean pLDDT, high uncertainty, low TM-score, and poor aligned coverage. This failure traces directly to the mechanism: with almost no homologous diversity, the MSA representation cannot provide reliable coevolutionary constraints, leaving the pair representation underconstrained for a hard FM target.

The strongest success, T1025-D1, shows the opposite regime. It has 13,053 MSA sequences and approximate Neff 8,521. Deep homologous support gives the model a rich statistical signal about compatible residue-residue relationships. This does not mean AlphaFold2 simply copies a template–this run used no templates–but it does mean that the Evoformer receives strong evolutionary evidence for the pair geometry that the structure module later realizes in 3D.

MSA depth is not deterministic. T1039-D1 has only 7 MSA sequences and approximate Neff 4.0, yet reaches TM-score 0.825. This matters because the model also uses single-sequence features and learned structural priors. Low Neff should therefore be interpreted as a warning sign, not as proof of failure. This matches the formal setup in Section 3.2: MSA features are inputs to the amortized predictor, but the benchmark loss is measured only after the full learned map, structure module, and recycling have acted.

5.3 Confidence and local calibration

Mean pLDDT is broadly useful, but it should not be read as a certificate of global correctness. It is nevertheless the strongest aggregate predictor in this benchmark: mean pLDDT and rank-1 TM-score have Spearman $\rho = 0.56$ with $p = 0.00118$. Low mean pLDDT correctly identifies the T1043-D1 failure, and high mean pLDDT usually accompanies accurate structures. The important exception is T1029-D1, which has mean pLDDT 94.95 but TM-score only 0.529. This is not a contradiction in the definition of pLDDT from Section 3.3: the confidence head estimates local reliability, not direct optimization of reference TM-score. A model can be locally confident in secondary-structure elements while the global reference alignment remains weak.

The residue-level analysis evaluates pLDDT on its natural scale. Across the pooled local-error table, higher pLDDT bins have lower median aligned C_α error. The absolute values in the right panel of Figure 4 are large because the table pools early recycle states, all five models, weak targets,

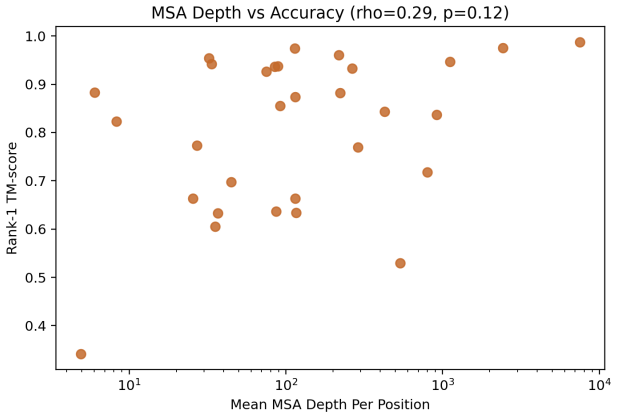


Figure 3: Approximate Neff versus rank-1 TM-score. T1043-D1 combines minimal MSA support with benchmark failure, while T1039-D1 shows that low MSA depth is not always fatal.

failure cases, final predictions, and poorly aligned residues. Therefore, the bar heights should not be read as median final rank-1 local errors. The important signal is directional: higher predicted local confidence corresponds to lower local structural error under broad sampling.

To check that this result is not only an artifact of pooling early recycle states, we repeated the binning on final rank-1 predictions only (Table 2). The restricted analysis is noisier: the intermediate bins are not monotone, which is expected because aligned C_α error after one global superposition is sensitive to domain shifts and target-level failures. The high-confidence boundary still carries signal: residues with pLDDT ≥ 90 have median aligned error 15.1 Å versus 22.2 Å for residues below 90. Treating individual residues as independent would overstate the evidence, so we also compared high- and lower-confidence median errors within targets. Among the 25 final rank-1 targets with residues in both groups, the high-confidence median is lower in 17 targets; the paired Wilcoxon test is directional but modest (one-sided $p = 0.048$, two-sided $p = 0.096$). This supports using pLDDT as a local reliability cue while also showing why local confidence cannot replace global structural evaluation.

pLDDT bin	Residues	Median C_α error (Å)
< 50	357	22.69
50–70	676	21.40
70–90	1,627	22.26
≥ 90	2,514	15.12

Table 2: Final rank-1 local calibration. The analysis is restricted to final rank-1 structures, rather than all models and recycle snapshots.

5.4 Model ranking reliability

ColabFold returns five models and ranks monomer predictions by confidence. In this benchmark, rank-1 is the best-by-TM-score model for 9/30 targets and is not best-by-TM-score for 21/30 targets. That count is less important than the effect size. The median rank-1 to best-of-five TM-score gap is only 0.003 and the mean gap is 0.014, so most strict mismatches are practically negligible. The scientifically interesting case is the tail: the maximum gap is 0.099 for T1036s1-D1, where inspecting alternative models would materially improve the reference score.

Mechanistically, this result is expected. The ranking score is a confidence proxy, not direct access to the unknown reference-based benchmark objective. If two models have similar local confidence, the confidence ranking may not select the model with the best global reference alignment. Thus, rank-1 is a reasonable default for many uses, but it should not be treated as a proof that alternative generated models are irrelevant.

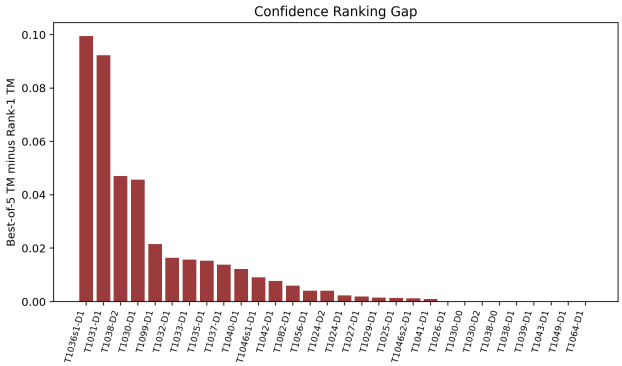
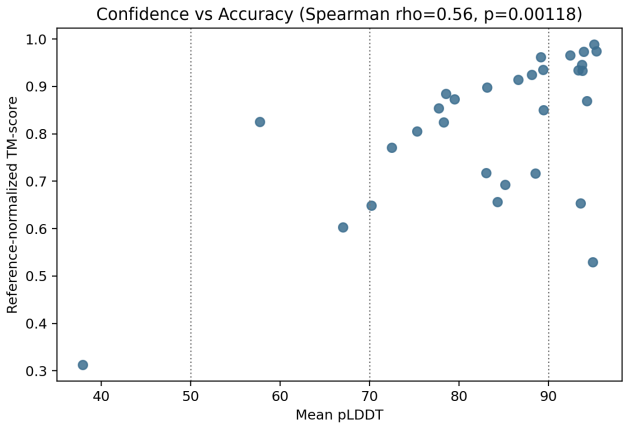


Figure 5: Gap between the top-ranked model and the best-of-five model by reference TM-score. The median gap is tiny, but a small number of targets have meaningful missed opportunity.

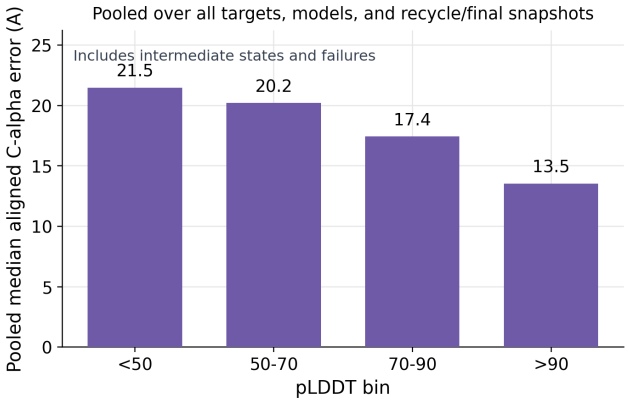
5.5 Recycling dynamics

Recycling provides the most direct observable trace of AlphaFold2's refinement mechanism. Across all targets and models, mean TM-score improves from 0.7047 at recycle 0 to 0.7687 at recycle 3, with intermediate means of 0.7392 and 0.7565. On average, recycling therefore moves the generated structures toward the reference. A paired target-level test supports that this is not just a descriptive fluctuation: after averaging the five models within each target, the recycle-3 scores are higher than recycle-0 scores for 23/30 targets (Wilcoxon signed-rank $W = 412$, one-sided $p = 3.96 \times 10^{-5}$; two-sided $p = 7.91 \times 10^{-5}$).

The average improvement does not mean recycling is a monotone optimizer of true structural accuracy. Under the rank-1 failure-mode definition, five targets regress across recycling and four show increasing confidence without accuracy gain. This follows from the structure-module and recycling formulation in Section 3.3: recycling iterates a learned map, not a guaranteed descent step on reference TM-score or physical free energy. The snapshots are useful because they expose stabilization, plateauing, or regression, but they should not be interpreted as a physical folding pathway.



(a) Mean pLDDT versus rank-1 TM-score.



(b) Pooled median C_α error by pLDDT bin.

Figure 4: Confidence calibration. Mean pLDDT is broadly informative (Spearman $\rho = 0.56$, $p = 0.00118$), but T1029-D1 shows that high local confidence does not guarantee strong global reference agreement. The pooled local-error panel combines targets, all five models, recycle snapshots, and final snapshots; the key result is the decrease in error with increasing pLDDT, not the absolute bar height.

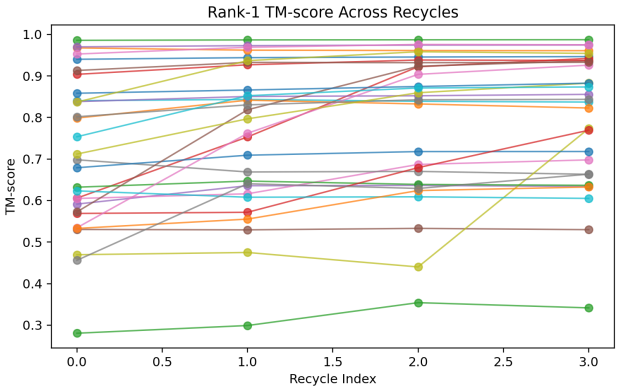


Figure 6: TM-score across recycle iterations. Mean accuracy improves, but individual trajectories are heterogeneous and not guaranteed to improve monotonically.

Flag	Count	Interpretation
Low TM-score flag	1	clear global fold failure
Mean pLDDT < 70	6	model self-reports uncertainty
High PAE	3	relative placement uncertainty
Rank-1 gap > 0	21	median gap only 0.003 TM-score
Recycle regression	5	final recycle lowers TM-score
Conf. up, acc. down	4	pLDDT rises without accuracy gain

Table 3: Representative failure-mode flags. Counts are not mutually exclusive.

5.6 Failure modes and runtime

The failure flags combine structural accuracy, confidence, ranking, and trajectory signals. They are not mutually exclusive because a target can be uncertain in several ways at once. The benchmark contains one low-TM-score global failure, six targets with mean pLDDT below 70, three with high PAE, twenty-one with a nonzero rank-1 to best-of-five gap, five with recycle regression, and four where confidence increases without accuracy gain. The rank gap count is deliberately paired with the effect size in Table 3, since most gaps are too small to matter biologically.

Runtime was modest for this scale. The full trajectory inference run completed in roughly 32 minutes on two RTX 6000 Ada GPUs. Runtime depends strongly on length, with the 621-residue T1036s1-D1 target taking 341 seconds. Trajectory instrumentation is therefore practical for a small benchmark, but would need prioritization for thousands of targets.

6 Case Studies

The aggregate results are best understood through three targets: a high-support success, a low-support failure, and a high-confidence mismatch. These cases connect empirical outcomes to the ColabFold inference map: MSA evidence shapes the learned representations, recycling refines but does not guarantee improvement, and confidence heads estimate model uncertainty but do not directly optimize reference TM-score.

Target	Neff	Final TM-score	RMSD	pLDDT
T1025-D1	8521.4	0.989	0.70	95.1
T1043-D1	2.0	0.313	5.05	37.9
T1029-D1	166.9	0.529	4.16	95.0

Table 4: Case-study final rank-1 metrics. RMSD is in Å.

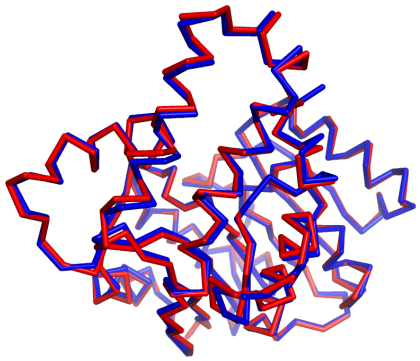


Figure 7: T1025-D1 structure overlay. Blue is the ColabFold prediction and red is the processed experimental reference.

T1025-D1: high-support success T1025-D1 is the cleanest example of AlphaFold2 working in the regime for which its mechanism is most favorable. It is a 257-residue TBM-easy target with 13,053 MSA sequences and approximate Neff 8,521. The final rank-1 prediction reaches TM-score 0.989 and RMSD 0.70 Å, with mean pLDDT 95.1. In the rank-1 trajectory, TM-score is already 0.986 at recycle 0 and remains 0.987 through the later recycle endpoints, while pLDDT rises from 93.2 to 95.4 and mean PAE decreases from about 3.8 to 3.2 Å.

The success follows from first principles of the model. A deep and diverse MSA gives the Evoformer strong evidence for residue couplings. Those couplings constrain the pair representation, reducing ambiguity in which residues should be close in 3D. The structure module then produces a geometry already near the reference, so recycling mainly sharpens confidence and small coordinate details. All diagnostics agree: deep MSA, low PAE, high pLDDT, stable recycling, full-domain alignment, and near-perfect overlay.

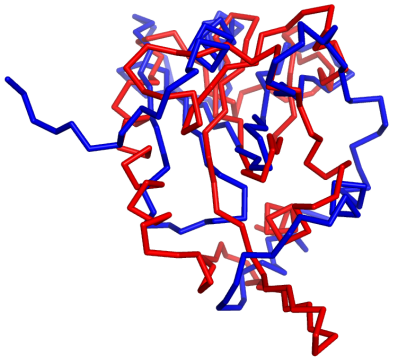


Figure 8: T1043-D1 structure overlay. Blue is the ColabFold prediction and red is the processed experimental reference.

T1043-D1: low-support failure T1043-D1 is the opposite regime. It is an FM target with only 6 MSA sequences and approximate Neff 2.0. The final rank-1 benchmark result has TM-score 0.313, RMSD 5.05 Å, mean pLDDT 37.9, and only 82/148 residues aligned. Across all five models, mean recycle TM-score barely changes from 0.284 at recycle 0 to 0.286 at recycle 3. In the rank-1 trajectory, the prediction reaches only moderate local improvements and remains globally wrong; mean PAE stays high, around 18–19 Å.

This failure follows directly from the model's information bottleneck. Free-modeling targets require long-range structural inference without a close template. With almost no homologous diversity, the MSA provides little covariation signal, so the pair representation is underconstrained. Recycling cannot recover missing information; it only reprocesses the model's current representations and coordinates. The low pLDDT is therefore an honest warning rather than a misleading confidence signal. Even without a reference, a user would see shallow MSA support, low local confidence, high aligned-error uncertainty, and inconsistent structure as reasons not to trust the output.

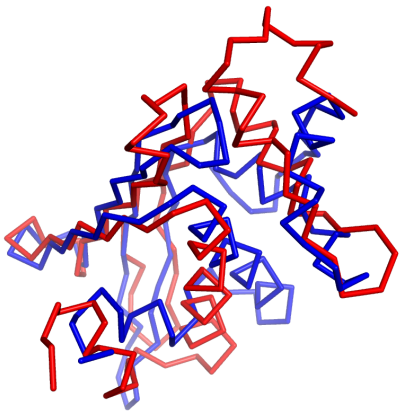


Figure 9: T1029-D1 structure overlay. Blue is the ColabFold prediction and red is the processed experimental reference.

T1029-D1: high-confidence mismatch T1029-D1 is the most important cautionary case. It is an FM target with approximate Neff 166.9, so it is not a no-MSA failure like T1043-D1. The final rank-1 prediction has mean pLDDT about 95.0 and mean PAE about 3.2 Å, but reference-normalized TM-score is only 0.529 and RMSD is 4.16 Å. The recycle trajectory does not resolve the issue: all-model mean TM-score is 0.534 at recycle 0 and 0.531 at recycle 3, while confidence increases modestly.

The first-principles explanation is that pLDDT is local. It asks whether the model expects local residue neighborhoods to be reliable, not whether the entire prediction optimizes reference-normalized TM-score. This benchmark does not establish that the prediction is an alternative biological conformer; that would require comparison to additional experimental structures or independent evidence of flexibility. The defensible conclusion is simpler and follows directly from the model objective: a prediction can be locally well formed and highly confident while still having weak global reference agreement. This case therefore marks the boundary of a common misuse: high pLDDT should support local claims about well-structured regions, but it should not by itself justify strong claims about global topology or functional mechanism.

7 Conclusion

This project evaluates ColabFold on 30 CASP14-derived targets and reframes a standard final-accuracy benchmark as a trustworthiness analysis. The final rank-1 predictions are strong: 29/30 targets reach topology-level success, with median TM-score 0.862. But the central conclusion is about interpretation. AlphaFold2-style predictions should not be treated as experimental structures. They are outputs of an amortized, confidence-scored generator that uses sequence, MSA evidence, learned structural priors, multiple model parameterizations, and recycling.

The diagnostics are informative precisely because they correspond to the formal model in Section 3.2. The optimization view in Section 3.2 explains why no single confidence number can replace reference-based structural loss. MSA support enters as input evidence for the learned predictor, but the non-significant MSA-depth correlation in this small benchmark shows that it should be interpreted as a warning signal, not a stand-alone accuracy law. The confidence definitions in Section 3.3 explain why pLDDT is the strongest aggregate predictor here while still failing to guarantee global TM-score for T1029-D1. The ranking rule in Section 3.3 explains why rank-1 need not maximize reference TM-score, even though the median gap is usually negligible. The recycling formulation in Section 3.3 explains why recycling can improve accuracy significantly on average without being a monotone optimizer. Therefore, the reliable use of AlphaFold2 and ColabFold requires reading the prediction as a structured model output: coordinates plus confidence, MSA support, model alternatives, and refinement behavior.

7.1 Limitations

This study is intentionally narrow. The benchmark contains 30 curated CASP14- derived targets rather than a large blind test set, so the aggregate numbers should be read as evidence about a controlled evaluation rather than a population-wide estimate of general performance. The inference runs also use ColabFold defaults without templates or Amber relaxation, so the results reflect one specific operational regime rather than every possible AlphaFold2- style setup. Finally, the trajectory analysis records discrete recycle endpoints and confidence outputs; it does not expose the full internal hidden state of the network. This study also does not directly compare these results with the official CASP14 assessor tables. Because the targets were reprocessed into a local benchmark and run under a specific no-template ColabFold setting, official CASP comparisons would require careful target/domain matching and method-condition alignment.

7.2 Future Work

Several extensions would make the analysis more complete. A larger benchmark would let the same trajectory-aware methodology probe a wider range of fold classes, domain architectures, and homology depths. It would also be useful to separate the effects of templates, relaxation, and alternative model selection policies, since those choices can materially change final accuracy and confidence calibration. A direct comparison to official CASP14 per-target assessments would also clarify where the trajectory-aware diagnostics add information beyond final model quality. More broadly, the same framework could be applied to other structure predictors or to protein design settings, where the question is not only whether a predicted structure is accurate, but whether the system's uncertainty estimates are trustworthy enough to guide downstream decision-making.

The broader lesson is that the useful unit of analysis is not just the final PDB file. It is the prediction record: the structure, confidence outputs, MSA support, ranked alternatives, and refinement trajectory considered together. That record gives users a more defensible basis for deciding when an

AlphaFold2-style prediction is strong evidence and when it should remain only a hypothesis.

References

Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; Bodenstein, S. W.; Evans, D. A.; Hung, C.-C.; O'Neill, M.; Reiman, D.; Tunyasuvunakool, K.; Wu, Z.; Zemgulyte, A.; Arvaniti, E.; Beattie, C.; et al. 2024. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630: 493–500.

Baek, M.; Anishchenko, I.; Humphreys, I. R.; Cong, Q.; Baker, D.; and DiMaio, F. 2023. Efficient and accurate prediction of protein structure using RoseTTAFold2. *bioRxiv*.

Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; Millán, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; van Dijk, A. A.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Garcia, K. C.; Grishin, N. V.; Adams, P. D.; Read, R. J.; and Baker, D. 2021. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557): 871–876.

Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; and Bourne, P. E. 2000. The Protein Data Bank. *Nucleic Acids Research*, 28(1): 235–242.

Bowie, J. U.; Luthy, R.; and Eisenberg, D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253(5016): 164–170.

Jones, D. T.; Buchan, D. W. A.; Cozzetto, D.; and Pontil, M. 2012. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, 28(2): 184–190.

Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; and Hassabis, D. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873): 583–589.

Kryshtafovych, A.; Schwede, T.; Topf, M.; Fidelis, K.; and Moult, J. 2021. Critical assessment of methods of protein structure prediction (CASP)—Round XIV. *Proteins: Structure, Function, and Bioinformatics*, 89(12): 1607–1617.

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

Mariani, V.; Biasini, M.; Barbato, A.; and Schwede, T. 2013. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21): 2722–2728.

Marks, D. S.; Colwell, L. J.; Sheridan, R.; Hopf, T. A.; Pagnani, A.; Zecchina, R.; and Sander, C. 2011. Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE*, 6(12): e28766.

Mirdita, M.; Schütze, K.; Moriwaki, Y.; Heo, L.; Ovchinnikov, S.; and Steinegger, M. 2022. ColabFold: making protein folding accessible to all. *Nature Methods*, 19(6): 679–682.

Rohl, C. A.; Strauss, C. E. M.; Misura, K. M. S.; and Baker, D. 2004. Protein structure prediction using Rosetta. *Methods in Enzymology*, 383: 66–93.

Sali, A.; and Blundell, T. L. 1993. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 234(3): 779–815.

Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Zidek, A.; Nelson, A. W. R.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan, S.; Kohli, P.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; and Hassabis, D. 2020. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792): 706–710.

Steinegger, M.; and Söding, J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11): 1026–1028.

Varadi, M.; Anyango, S.; Deshpande, M.; Nair, S.; Natassia, C.; Yordanova, G.; Yuan, D.; Stroe, O.; Wood, G.; Laydon, A.; Zidek, A.; Green, T.; Tunyasuvunakool, K.; Petersen, S.; Jumper, J.; Clancy, E.; Green, R.; Vora, A.; Lutfi, M.; Figurnov, M.; Cowie, A.; Hobbs, N.; Kohli, P.; Kleywegt, G.; Birney, E.; Hassabis, D.; and Velankar, S. 2022. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1): D439–D444.

Wu, R.; Ding, F.; Wang, R.; Shen, R.; Zhang, X.; Luo, S.; Su, C.; Wu, Z.; Xie, Q.; Berger, B.; Ma, J.; and Peng, J. 2022. High-resolution de novo structure prediction from primary sequence. *bioRxiv*.

Yang, J.; Anishchenko, I.; Park, H.; Peng, Z.; Ovchinnikov, S.; and Baker, D. 2020. Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences*, 117(3): 1496–1503.

Yang, J.; Yan, R.; Roy, A.; Xu, D.; Poisson, J.; and Zhang, Y. 2015. The I-TASSER Suite: protein structure and function prediction. *Nature Methods*, 12(1): 7–8.

Zhang, C.; Shine, M.; Pyle, A. M.; and Zhang, Y. 2022. US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nature Methods*, 19(9): 1109–1115.

Zhang, Y.; and Skolnick, J. 2004. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4): 702–710.

A Full Per-Target Results

Target	Cat.	Len.	TM-score	RMSD	pLDDT	Neff
T1024-D1	TBM-easy	193	0.946	1.32	93.8	632.6
T1024-D2	TBM-easy	204	0.962	1.18	89.2	155.7
T1025-D1	TBM-easy	257	0.989	0.70	95.1	8521.4
T1026-D1	TBM-hard	146	0.898	1.55	83.2	8.1
T1029-D1	FM	125	0.529	4.16	95.0	166.9
T1030-D1	TBM-hard	154	0.653	3.94	93.6	84.5
T1030-D2	TBM-hard	119	0.805	1.98	75.3	3.0
T1031-D1	FM	95	0.914	1.49	86.6	380.8
T1032-D1	TBM-hard	170	0.717	2.84	88.5	36.4
T1033-D1	FM	100	0.824	2.47	78.3	8.0
T1035-D1	FM/TBM	102	0.771	2.41	72.5	68.6
T1038-D1	FM	114	0.925	1.48	88.2	13.7
T1038-D2	FM/TBM	76	0.850	1.91	89.5	62.1
T1039-D1	FM	161	0.825	2.24	57.8	4.0
T1040-D1	FM	130	0.603	3.20	67.0	38.1
T1041-D1	FM	242	0.885	2.41	78.5	295.2
T1043-D1	FM	148	0.313	5.05	37.9	2.0
T1046s1-D1	FM/TBM	72	0.933	0.93	93.8	36.6
T1046s2-D1	TBM-hard	141	0.966	0.88	92.4	42.3
T1049-D1	FM	134	0.934	1.19	93.4	120.2
T1056-D1	TBM-hard	169	0.974	0.88	95.4	2646.5
T1064-D1	FM	92	0.873	1.89	79.5	4.9
T1082-D1	FM/TBM	75	0.973	0.53	94.0	6.0
T1099-D1	TBM-hard	178	0.869	2.18	94.3	14.7
T1027-D1	FM	99	0.656	2.55	84.3	40.1
T1036s1-D1	TBM-easy	621	0.718	5.18	83.0	150.0
T1037-D1	FM	404	0.854	3.01	77.7	168.9
T1042-D1	FM	276	0.649	4.56	70.2	50.7
T1030-D0	MultiDom	273	0.692	3.99	85.2	40.5
T1038-D0	MultiDom	190	0.935	1.70	89.4	15.3

Table 5: Per-target final rank-1 results. The appendix table is split into two side-by-side halves for readability, but the content is unchanged. Length is the benchmark domain length. RMSD is in Å. Neff is the approximate diversity-weighted MSA depth.