DOES REASONING IMPROVE LLMS BEYOND MATHEMATICS? A BIOLOGICAL TEST ON INTRINSICALLY DISORDERED PROTEIN REGIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reasoning improves language models in domains such as mathematics, where premises and answers are well defined. Whether these benefits extend to biology, where evidence is incomplete and heterogeneous, remains unclear. We study this question through prediction of intrinsically disordered protein regions (IDRs). GPT-5.4 generated evidence-grounded teacher traces that were transferred to Qwen3-4B through supervised fine-tuning (SFT) and Group Relative Policy Optimization (GRPO). We combined the resulting models with a frozen ESM-2 residue classifier and calibrated interval decoder. Zero-shot Qwen produced valid structured outputs but no usable IDR intervals, showing that formatting alone does not confer biological prediction ability. Both hybrid models performed strongly on teacher-corpus validation and retained substantial performance on an independently selected, sequence-disjoint test set, numerically exceeding two leading CAID2 Disorder-NOX predictors evaluated with the same labels. GRPO improved interval localization, trace–answer consistency, and reasoning coherence, but did not consistently improve residue classification over SFT. Perturbation analysis identified candidate-region evidence as the dominant predictive signal, while claim-level auditing showed that coherent traces could still contain unsupported statements. Overall, our reasoning-based framework improves selected biological prediction objectives under incomplete evidence, but its benefits are objective-dependent; the strongest results arise from combining language-model reasoning with specialized protein representations and explicit grounding checks.

1 INTRODUCTION

Explicit reasoning has improved language-model performance in mathematics and symbolic problem solving, where premises, operations, and correct answers are usually well defined and readily verified (Wei et al., 2022; Kojima et al., 2022; Lightman et al., 2024). It is less clear whether these benefits extend to biology. Biological inference often requires combining heterogeneous evidence of uneven quality, while relevant measurements may be incomplete and the relationship between evidence and conclusion is less deterministic. We therefore ask whether reasoning can improve an LLM's ability to infer intrinsically disordered protein regions (IDRs).

IDRs are protein segments that do not adopt a single stable three-dimensional structure under physiological conditions. Their conformational flexibility supports signaling, molecular recognition, regulation, phase separation, and protein-complex assembly, while their dysregulation is associated with cancer and neurodegenerative disease (Wright & Dyson, 2015; Babu et al., 2011; Banani et al., 2017; Holehouse & Kragelund, 2023). Because IDRs form dynamic, context-dependent ensembles, they are difficult to characterize experimentally, and available annotations remain incomplete.

Most computational IDR predictors formulate the task as residue-level classification. These methods can accurately estimate disorder propensity but generally do not explain how sequence, structural, and evolutionary evidence supports a prediction. LLMs offer a complementary capability: they can integrate heterogeneous evidence, generate structured reasoning traces, and return protein-level and interval-level predictions. However, explicit reasoning may not improve every aspect of IDR prediction, particularly because generative boundary selection and dense residue classification im-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

pose different modeling requirements. Reasoning quality must also be evaluated separately from predictive accuracy, since a coherent explanation may still contain unsupported biological claims.

We investigate these questions using an evidence-guided IDR prediction framework. GPT-5.4 generated structured teacher traces and IDR answers, which were converted into a five-step boundary-aware format termed Evidence-Grounded Boundary SFT (EGB-SFT). We used these data to fine-tune Qwen models through parameter-efficient adaptation (Yang et al., 2025; Hu et al., 2022). The selected Qwen3-4B model was combined with an auxiliary residue classifier operating on frozen ESM-2 representations and a calibrated interval decoder (Lin et al., 2023). We then applied Group Relative Policy Optimization (GRPO) to the SFT-initialized generative branch. This design enables us to test whether reasoning and reinforcement learning improve interval inference, whether specialized protein representations remain necessary for residue-level prediction, and whether generated explanations remain grounded in the available evidence.

Contributions. Our main contributions are:

- We construct an evidence-grounded IDR knowledge base containing protein sequences, heterogeneous biological evidence, structured teacher reasoning traces, and residue-interval answers.
- We develop Qwen3-4B SFT and GRPO hybrid models that combine biological reasoning with frozen ESM-2 residue representations and normalized interval decoding. Both models generalize to an independently selected, sequence-disjoint test set and numerically outperform two leading CAID2 Disorder-NOX predictors evaluated with the same labels (Wang et al., 2024; Kabir & Hoque, 2024).
- We introduce a multi-layer evaluation framework that measures structured output validity, predictive agreement, reasoning coherence, and claim-level evidence grounding.
- We use perturbation and reward-ablation studies to characterize how sequence, structural, evolutionary, and reasoning components contribute to IDR inference.

2 Related Work

Reasoning in well-specified domains. Explicit reasoning has improved language-model performance in mathematics and symbolic problem solving, where premises, operations, and correct answers are typically well defined. Chain-of-thought prompting elicits intermediate reasoning before the final answer (Wei et al., 2022), while zero-shot chain-of-thought shows that prompting the same model to reason step by step can improve accuracy (Kojima et al., 2022). Self-consistency aggregates multiple reasoning trajectories (Wang et al., 2023), and process supervision evaluates intermediate steps rather than only final answers (Lightman et al., 2024). These results establish the value of explicit reasoning when correctness can be evaluated against relatively clear rules.

Reasoning language models in biology. Recent work has extended structured reasoning to biological tasks. ProLLM represents signaling pathways as intermediate reasoning paths for protein–protein interaction prediction (Jin et al., 2024). ProteinReasoner uses evolutionary profiles as an intermediate representation between protein sequence and structure for multimodal prediction and design (Liu et al., 2025). CellDuality applies reinforcement learning and bidirectional self-verification to single-cell reasoning (Chen et al., 2026). BioReason-Pro combines protein representations, biological context, teacher-generated traces, and reinforcement learning for protein-function prediction (Fallahpour et al., 2026). These studies demonstrate growing interest in biological reasoning, but whether explicit reasoning consistently improves residue-level prediction remains unresolved.

Intrinsic-disorder prediction and CAID. IDR predictors traditionally use amino-acid composition, physicochemical properties, evolutionary profiles, predicted structural features, and learned sequence representations to estimate disorder at residue resolution. The Critical Assessment of Protein Intrinsic Disorder Prediction (CAID) provides blind community benchmarks based primarily on experimentally supported DisProt annotations (Necci et al., 2021; Mehdiabadi et al., 2026). CAID evaluates predictors using AUROC, average precision, maximum F1, Matthews correlation coefficient, and prediction coverage. Its alternative reference sets also show that measured performance

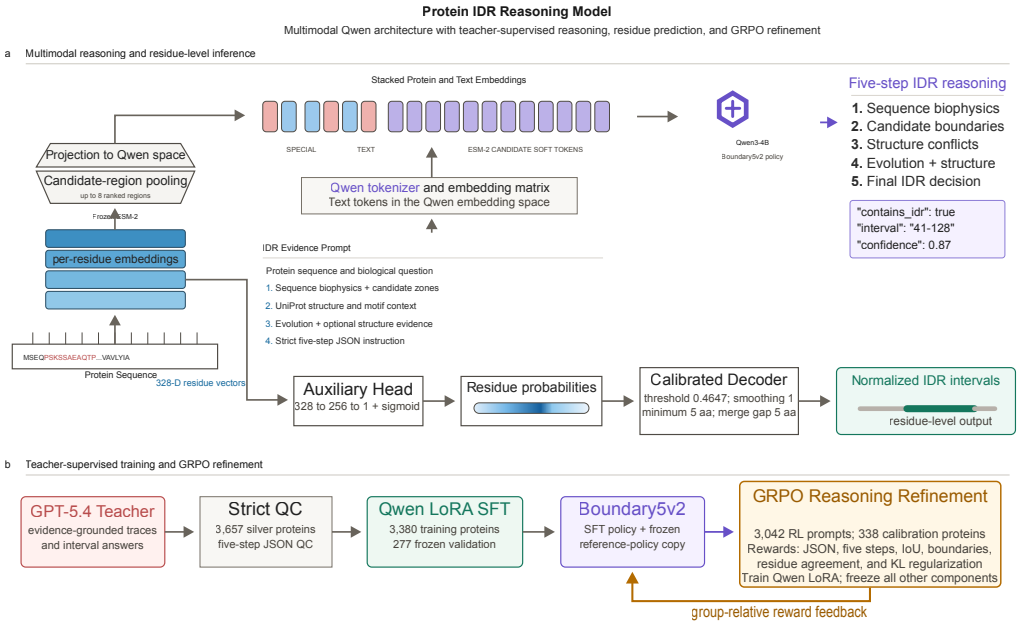


Figure 1: Architecture of the protein IDR reasoning framework. Qwen3-4B was trained using SFT and GRPO, while a frozen ESM-2 auxiliary classifier supplied dense residue probabilities. Generated intervals and auxiliary scores were fused before normalized decoding.

depends on the definitions of positive, negative, and uncertain residues. Recent CAID methods increasingly use deep learning and protein language-model representations, but most remain discriminative predictors that return residue scores without an accompanying biological rationale.

IDR annotations and biological evidence. DisProt provides manually curated, experimentally supported disorder annotations, whereas MobiDB integrates experimental annotations with consensus and computational predictions (Aspromonte et al., 2024; Piovesan et al., 2025). Sequence composition, structural annotations, evolutionary conservation, and interaction context provide complementary evidence for disorder. AlphaFold confidence can offer additional support, although low pLDDT may reflect flexibility, limited evolutionary information, or modeling difficulty and is not direct proof of intrinsic disorder (Jumper et al., 2021; Ruff & Pappu, 2021). Studies of IDR evolution and conformational behavior further demonstrate that disorder cannot always be inferred reliably from a single evidence source (Singleton & Eisen, 2024; Tesei et al., 2024).

3 Experiments

Figure 1 summarizes the framework. We constructed an IDR knowledge base, generated GPT-5.4 teacher traces, fine-tuned Qwen models, and refined the selected Qwen3-4B checkpoint using GRPO. SFT- or GRPO-generated intervals were then fused with probabilities from a frozen ESM-2 auxiliary residue classifier and converted into normalized intervals by a calibrated decoder. We compared both hybrid models with zero-shot Qwen3-4B, flDPnn2, and DisPredict3.0 on locked teacher-corpus validation and independent test cohorts.

3.1 Data, Evidence, and Evaluation Cohorts

We assembled 20,431 reviewed human proteins from UniProtKB, including canonical sequences, stable accessions, gene and protein names, organism metadata, and residue-level annotations (The UniProt Consortium, 2023). These annotations included domains, functional regions, repeats, transmembrane helices, signal peptides, and motifs. Reference IDR intervals from DisProt and MobiDB were harmonized by standardizing coordinates and field formats (Aspromonte et al., 2024; Piovesan

et al., 2025). Source database and evidence category were retained for provenance and quality control but withheld from GPT-5.4 and Qwen to prevent target-source leakage.

Sequence-derived evidence was calculated in sliding windows of 9, 15, 21, 41, and 81 residues. Features included hydropathy, net charge per residue, charged- and polar-residue fractions, aromatic content, Pro/Gly/Ser/Gln enrichment, low complexity, and a sequence-derived disorder-tendency score. Overlapping disorder-compatible windows were merged into candidate zones. Candidate zones were then intersected with UniProt annotations: domains, transmembrane segments, and signal peptides were treated as potential structural conflicts, whereas repeats, functional regions, and motifs were retained as contextual evidence.

Gene-Ages supplied whole-gene evolutionary age and uncertainty, while Ensembl Compara and OMA supplied orthology information (Liebeskind et al., 2016; Herrero et al., 2016; Altenhoff et al., 2024). Where regional alignments were available, we calculated ortholog presence, alignment coverage, gap frequency, compositional similarity, and the oldest supported evolutionary level. Regional ortholog features were available for 247 of the final 3,657 teacher-corpus proteins. Whole-gene age was treated as contextual evidence and not as the age of an individual IDR.

AlphaFold pLDDT was treated as optional supporting evidence: low pLDDT was considered compatible with flexibility but was not sufficient to establish intrinsic disorder (Jumper et al., 2021; Varadi et al., 2024). Regional pLDDT values were absent from the final teacher prompts and did not affect corpus selection. ESM-2 representations were computed locally from the protein sequences (Lin et al., 2023). Candidate-region embeddings were used as Qwen soft-token inputs, while full-length residue embeddings were retained for the auxiliary residue classifier.

Evidence and reference annotations were stored in separate fields, and only permitted evidence was serialized into model prompts. Proteins were ranked using sequence support, candidate/reference agreement, available regional evolutionary evidence, and structural-conflict penalties. Candidate zones were intentionally overcomplete and included nonmatching regions that could be rejected or trimmed during supervision. This procedure produced an evidence-supported teacher corpus of 3,657 proteins, divided into 3,380 training and 277 accession-disjoint validation proteins.

To evaluate dependence on this construction procedure, we selected 277 held-out proteins from the broader knowledge base. Teacher-corpus accessions and identical sequences were excluded, and neither candidate zones nor candidate/reference overlap was used to determine membership. The locked independent cohort was used only for evaluation and not for model training, checkpoint selection, fusion-weight selection, or decoder calibration. We also constructed a sequence-disjoint pool of 3,380 proteins for future independent training experiments; that pool was not used in any result reported here.

3.2 Teacher Traces and Supervised Fine-Tuning

GPT-5.4 generated an evidence-grounded explanation for each supplied reference answer. It received the permitted protein evidence together with the target IDR intervals and confidence values and was instructed to explain, rather than independently infer, the supplied answer without naming hidden annotation sources. Compact teacher prompts contained sequence-derived summaries rather than the literal amino-acid string.

Each teacher trace covered six operations: sequence-biophysics assessment, candidate-zone boundary assessment, structural-conflict assessment, AlphaFold pLDDT assessment, evolutionary assessment, and final boundary decision. The answer reported IDR presence, interval count, boundaries, and confidence values. All 3,657 records passed structural quality control, including unique accessions, six nonempty reasoning steps, and complete answers. Together, they contained 21,942 step-level observations without heuristic evidence of truncation.

For SFT, the teacher outputs were compressed into the five-step EGB-SFT format: sequence biophysics, candidate-zone and boundary assessment, structural conflicts, combined AlphaFold and evolutionary assessment, and final boundary selection. Each input contained the sequence-derived evidence and up to eight candidate-region soft tokens constructed from pooled ESM-2 embeddings and regional features. Candidate zones received retain, trim, or reject decisions, with boundaries supervised during both candidate assessment and final selection. The target was a single JSON object containing the reasoning trace and structured IDR answer.

We compared Qwen2.5-3B-Instruct, Qwen3-4B-Instruct-2507, and Qwen3-8B using identical partitions, inputs, supervision targets, and 1,200 optimization steps (Yang et al., 2024; 2025). LoRA was applied to the query, key, value, and output projections with rank 4, scaling factor 8, and dropout 0.05 (Hu et al., 2022). Training used a learning rate of 2×10^{-4}, batch size 1, and four gradient-accumulation steps. Qwen3-4B achieved the strongest development performance and was selected for subsequent experiments. Its final run used a 2,048-token context and generated up to 1,200 new tokens.

3.3 GRPO Refinement

Starting from the selected Qwen3-4B SFT checkpoint, we applied Group Relative Policy Optimization using the same biological evidence and candidate-region soft tokens. Multiple completions were sampled for each prompt and scored relative to other completions in the same group. The reward combined answer accuracy, structured-output validity, and trace–answer consistency:

$$R = R_{\text{answer}} + 0.2R_{\text{format}} + 0.1R_{\text{consistency}}.$$

The answer term incorporated residue agreement, interval overlap, and IDR-count accuracy. The formatting term evaluated valid JSON, presence of the required reasoning operations and final answer, accession consistency, and absence of prompt leakage. The consistency term evaluated agreement between intermediate candidate decisions, numerical claims, and final intervals.

We evaluated learning rates $\{5 \times 10^{-6}, 2.5 \times 10^{-5}, 10^{-4}\}$ and group sizes $\{8, 16\}$ for up to 1,690 optimization steps. Checkpoints were selected using development residue F1, mean interval IoU, and reward stability rather than reward alone (Gao et al., 2023). The selected run used a learning rate of 2.5×10^{-5} and group size 8. The SFT checkpoint served as the reference policy, and the independent test set was not used for hyperparameter or checkpoint selection.

3.4 Hybrid Residue Prediction

Because generative Qwen predicts discrete intervals rather than dense residue scores, we trained a two-layer classifier on frozen residue-level ESM-2 embeddings while leaving Qwen unchanged. The classifier produced one disorder probability per residue. The 3,657 teacher-corpus proteins were divided into 3,042 auxiliary-head training proteins, 338 calibration proteins, and the locked 277-protein validation set.

The calibration subset was used to select the auxiliary-head checkpoint and a deterministic interval decoder. The frozen decoder used a probability threshold of 0.4647, smoothing width of one residue, minimum interval length of five residues, and maximum merge gap of five residues. Neither locked evaluation cohort was used for these choices.

For both SFT and GRPO, generated intervals were converted into a residue mask and fused with the auxiliary probabilities:

$$p_{\text{fused}}(i) = \sigma[\text{logit}(p_{\text{aux}}(i)) + \alpha m(i)],$$

where $m(i) = +1$ inside a generated interval and -1 otherwise. We prespecified $\alpha = 0.5$ without tuning it on either evaluation cohort and applied the same frozen decoder to SFT and GRPO. Malformed, interval-missing, or accession-mismatched generations reverted to auxiliary-only scores. Fusion used only structured interval proposals and did not inspect or modify the reasoning text; the original traces were retained for grounding evaluation.

3.5 Baselines and Predictive Evaluation

Zero-shot Qwen3-4B-Instruct-2507 was evaluated without SFT, adapters, an auxiliary classifier, fusion, or interval normalization. Prompts contained either the sequence alone or the sequence with the biological evidence available to the trained model. Reference intervals, teacher traces, and annotation provenance were withheld, and deterministic decoding used the same five-step JSON format as the trained system.

We also evaluated flDPnn2 and DisPredict3.0, which led the CAID2 Disorder-NOX benchmark in average precision and AUROC, respectively (Wang et al., 2024; Kabir & Hoque, 2024). Disorder-NOX was selected because its residue-label convention was closer to ours than the structure-restricted Disorder-PDB task. For matched-cohort evaluation, all methods used the same frozen

labels: residues within reference IDRs were positive and all other scored residues were negative. This harmonized the evaluation labels, although the predictors used different training data and our labels did not reproduce the full CAID exclusion mask.

Continuous residue scores were pooled across proteins to calculate AUROC, AUPRC, $F_{\max}$, MCC at $F_{\max}$, and prediction coverage. Fixed-decoder predictions were evaluated using precision, recall, F1, MCC, specificity, accuracy, and balanced accuracy. For the Qwen systems, we additionally measured JSON validity, interval precision, recall, F1, mean best IoU, IDR-count accuracy, underprediction and overprediction, and boundary error normalized by protein length. Interval metrics were not computed for the external predictors because they were evaluated as residue-level methods. Both locked cohorts contained only IDR-positive proteins; therefore, protein-level specificity could not be estimated.

3.6 Reasoning-Trace Evaluation

We evaluated reasoning quality independently from predictive accuracy. Prompt-supported claims were extracted as evidence-region pairs and classified as supported, contradicted, or unverifiable. Claims corresponding to the initial reasoning operations were checked against evidence visible in the prompt. Final intervals were additionally compared with source-specific annotations from DisProt and MobiDB and with a low-pLDDT flexibility proxy in that priority order, using record-specific annotations or a pooled lookup spanning 3,416 proteins. Unverifiable claims were not counted as failures.

ROSCOE assessed internal trace coherence using step-level entailment, semantic similarity, chain faithfulness, and answer coverage (Golovneva et al., 2022). These scores were reported separately from factual grounding because a coherent trace can contain unsupported claims, while a correct answer can arise from incomplete intermediate reasoning.

GPT-6 Astra independently evaluated SFT and GRPO traces for the same 277-protein validation cohort. For each protein, the judge received only the evidence originally shown to Qwen and the generated trace and answer. Teacher traces, normalized references, and annotation provenance were withheld. The judge decomposed each output into atomic claims and distinguished supported statements, reasonable inferences, prediction decisions, confidence-calibration issues, boundary-justification gaps, unsupported biological rationales, fabricated or contradicted evidence, and unverifiable claims.

Protein-level judgments distinguished minor issues, substantive grounding issues, and fabricated or contradicted evidence. Duplicate semantic claims in structured fields and answer text were counted once. Confidence issues were analyzed separately because candidate-region scores were not calibrated probabilities of the final predictions. Reference intervals were introduced only after judging to calculate deterministic interval agreement and were never supplied to GPT-6 Astra.

3.7 Evidence Perturbation

We separately perturbed candidate-zone evidence, protein-header information, sequence biophysics, structural conflicts, and evolutionary evidence while holding all other inputs fixed. Inference-time perturbations evaluated which signals directly influenced the fitted policy. Training-time perturbations retrained GRPO from the same SFT initialization using the same learning rate, group size, and training budget, then evaluated each policy on clean development inputs.

Perturbations were first screened using the change in negative log-likelihood of the reference answer, with each token conditioned on the preceding reference tokens. Top-ranked conditions were then evaluated by full generation. Protein-level differences from the unperturbed condition were summarized using 95% bootstrap confidence intervals with 2,000 resamples and paired two-sided Wilcoxon signed-rank tests. Paired analyses were restricted to proteins whose outputs were successfully parsed in both conditions.

4 RESULTS

4.1 QWEN3-4B SFT HYBRID MODEL

The SFT hybrid combined Qwen3-4B interval proposals with frozen auxiliary-head scores using $\alpha = 0.5$ and the calibrated normalized decoder. On the teacher-corpus validation set, it achieved pooled AUPRC of 0.9848, residue F1 of 0.9474, and mean best interval IoU of 0.7830. Performance decreased on the independently selected test set but remained strong, with AUPRC of 0.8986, residue F1 of 0.8423, and mean best IoU of 0.5224. Structured SFT intervals were usable for 93 of 277 independent-test proteins (33.6%); the remaining proteins retained auxiliary-only scores.

4.2 QWEN3-4B GRPO HYBRID MODEL

The GRPO hybrid used the same auxiliary head, $\alpha = 0.5$ fusion, and decoder. It achieved AUPRC of 0.9851, residue F1 of 0.9463, and mean best IoU of 0.7795 on teacher-corpus validation. On the independent test, the corresponding values were 0.8928, 0.8377, and 0.5495.

The SFT and GRPO hybrids performed similarly on teacher-corpus validation. On the independent test, SFT produced slightly stronger residue-level classification, whereas GRPO increased interval F1 from 0.7399 to 0.7480 and mean best IoU from 0.5224 to 0.5495. Thus, GRPO improved independent-set interval localization but did not uniformly improve residue-level performance.

4.3 COMPARISON WITH BASELINE MODELS

Zero-shot Qwen3-4B produced valid JSON for 100.0% of sequence-only prompts and 99.3% of sequence-plus-evidence prompts but generated no usable IDR intervals, yielding residue F1 and mean best IoU of zero. Output-format compliance alone therefore did not confer IDR-prediction ability.

Both Qwen hybrid models also numerically outperformed the external flDPnn2 and DisPredict3.0 baselines on the same locked cohorts. On the independent test, the strongest external baseline, flDPnn2, achieved AUPRC of 0.7983 and $F_{\max}$ of 0.7535, compared with 0.8986 and 0.8387 for the SFT hybrid. These comparisons use identical evaluation labels but do not imply that the models used identical training data.

Detailed model results are provided in Appendix Table A.1.

4.4 EVIDENCE PERTURBATION ANALYSIS

Candidate-zone evidence produced the largest performance degradation at both inference and GRPO training time (Table 1). Removing it at inference reduced residue F1 by 0.2088 and mean IoU by 0.2580 ($p < 0.001$); perturbing it during GRPO training reduced the same metrics by 0.0565 and 0.0996 ($p < 0.001$). Protein-header information also contributed significantly at both stages. Structural-conflict evidence affected the fixed policy at inference but had no significant training-time effect.

Biophysical and evolutionary evidence showed weaker, stage-dependent effects. Neither significantly affected inference-time performance. During GRPO training, perturbing evolutionary age reduced residue F1 by 0.0178 ($p = 0.019$), whereas perturbing biophysical evidence produced a small increase in mean IoU ($\Delta = 0.0081$, $p = 0.021$). These results suggest that evidence used directly by a fitted policy can differ from evidence that influences policy formation during reinforcement learning.

4.5 GROUNDING AND COHERENCE OF GRPO-GENERATED REASONING TRACES

Relative to SFT, the selected GRPO checkpoint improved residue-level agreement with source-specific DisProt annotations (Jaccard 0.295 to 0.453) and with the low-pLDDT flexibility proxy (0.558 to 0.635). These comparisons are diagnostic rather than fully independent because DisProt and MobiDB contributed to construction of the reference corpus. ROSCOE step entailment increased from 0.0656 to 0.1574, and chain faithfulness increased from 0.0193 to 0.3415, indicating

Table 1: Effects of inference-time and GRPO training-time evidence perturbations on IDR prediction. Changes are measured relative to the corresponding unperturbed condition. Negative values indicate lower performance after perturbation; p values are from paired two-sided Wilcoxon signed-rank tests.

	Inference-time perturbation				GRPO training-time perturbation			
	Δ Residue F1		Δ Mean IoU		Δ Residue F1		Δ Mean IoU	
Condition	Δ	p	Δ	p	Δ	p	Δ	p
Candidate zones	$\mathbf{-0.2088}$	< 0.001	$\mathbf{-0.2580}$	< 0.001	$\mathbf{-0.0565}$	< 0.001	$\mathbf{-0.0996}$	< 0.001
Protein header	$\mathbf{-0.0607}$	< 0.001	$\mathbf{-0.0762}$	< 0.001	$\mathbf{-0.0257}$	0.016	$\mathbf{-0.0226}$	0.043
Structure conflicts	$\mathbf{-0.0554}$	< 0.001	$\mathbf{-0.0348}$	0.002	-0.0081	0.342	$+0.0097$	0.109
Biophysics	-0.0013	0.865	-0.0073	0.186	-0.0100	0.497	$\mathbf{+0.0081}$	0.021
Evolution age	-0.0032	0.580	-0.0012	0.989	$\mathbf{-0.0178}$	0.019	$+0.0020$	0.946

stronger internal coherence. Adding source-agreement verification to the reward at weight 0.1 produced modest gains in residue F1 and IoU, but no reward variant consistently improved all metrics. Detailed results are provided in Appendix Table B.2.

All 277 SFT and GRPO outputs were valid JSON. Under the strict GPT-6 Astra record-level rubric, no GRPO output was fully grounded. Among the GRPO traces, 206 proteins (74.4%) contained only minor issues, 64 (23.1%) contained a substantive grounding issue, and 7 (2.5%) contained fabricated or contradicted evidence. Thus, 71 GRPO outputs (25.6%) contained a substantive or critical issue, compared with 84 SFT outputs (30.3%). Trace–answer consistency increased from 193 proteins (69.7%) under SFT to 250 (90.3%) under GRPO (Appendix Table B.5).

At the claim level, 1,219 of 1,920 GRPO claims (63.5%) had no identified issue; 617 (32.1%) were minor, 76 (4.0%) were substantive, and 8 (0.4%) were critical. The identified issues included 277 confidence-calibration claims, 240 boundary-justification gaps, 173 unsupported biological rationales, 8 fabricated-evidence claims, and 3 contradictions. Compared with SFT, GRPO reduced substantive or critical claims and contradictions but increased boundary-justification gaps and unsupported biological rationales. Thus, GRPO improved trace–answer consistency and some dimensions of grounding without eliminating unsupported reasoning.

For the raw generative outputs used in the grounding audit, GRPO improved mean best IoU from 0.6399 to 0.6749 and increased the fraction of proteins with IoU ≥ 0.50 from 65.0% to 70.4%, although exact interval agreement decreased from 23.8% to 21.7%. Within the GRPO outputs, traces with only minor grounding issues had higher mean best IoU than those with substantive issues or fabricated or contradicted evidence (0.7340 versus 0.5143 and 0.4029, respectively). This association is descriptive and does not establish that grounding errors caused poorer interval predictions. Together, the ROSCOE and GPT-6 evaluations show that internal coherence and factual grounding are related but distinct properties of generated biological reasoning.

5 Discussion

The Qwen3-4B hybrid models achieved residue-level performance numerically above the two leading CAID2 Disorder-NOX baselines evaluated on the same locked cohorts. This supports combining language-model interval proposals with frozen protein representations for dense prediction. The comparison is informative but not definitive because the models used different training data and our evaluation did not reproduce the complete CAID protocol.

To our knowledge, this is the first study to combine human-readable, step-by-step LLM reasoning with residue- and interval-level IDR prediction and to demonstrate numerically stronger performance than leading CAID2 Disorder-NOX predictors evaluated on matched cohorts. Unlike methods that map protein representations directly to residue scores, our framework produces an auditable reasoning trace while integrating multiple evidence layers, including sequence biophysics, structural counterevidence, evolutionary context, and learned protein representations. These results suggest that explicit reasoning can contribute to a biological task that cannot be resolved by a single deterministic rule or evidence source.

Performance was slightly lower on the independently selected, sequence-disjoint cohort than on teacher-corpus validation, but remained in the same predictive regime and continued to exceed both external baselines. The main residue-level findings therefore generalized beyond the proteins used to construct the teacher corpus. The independent cohort provides a conservative estimate because it was not used for training, calibration, checkpoint selection, fusion-weight selection, or decoder tuning. Nevertheless, statistical equivalence between the two cohorts was not established.

GRPO improved several aspects of SFT. On the independent test, interval F1 increased from 0.7399 to 0.7480 and mean best IoU from 0.5224 to 0.5495. Trace–answer consistency increased from 69.7% to 90.3%, while the proportion of proteins with substantive or critical grounding issues decreased from 30.3% to 25.6%. These gains did not extend uniformly to residue classification, for which SFT remained slightly stronger. GRPO therefore primarily improved interval localization and reasoning consistency rather than every predictive objective.

Perturbation and reward ablations help explain these results. Candidate-region evidence caused the largest losses when perturbed at inference or during GRPO training, while protein-header information contributed at both stages. Structural-conflict evidence was particularly important to the fitted inference policy, whereas biophysical and evolutionary evidence had smaller, stage-dependent effects. Reward modifications produced modest and nonuniform changes: adding source-agreement verification at weight 0.1 improved residue F1 and interval IoU, but the default reward remained strongest on several other metrics. These findings indicate that performance depends on integrating complementary biological evidence rather than emphasizing a single feature or reward component.

The GPT-6 Astra audit provides a nuanced view of reasoning quality. All GRPO outputs were valid JSON, and 74.4% of proteins contained only minor grounding issues. At the claim level, 63.5% of statements had no identified issue and only 4.4% were classified as substantive or critical. Fabricated or contradicted evidence occurred in 2.5% of proteins and 0.58% of claims. Relative to SFT, GRPO reduced contradictions and substantive or critical protein-level judgments, although it increased boundary-justification gaps and unsupported biological rationales. GRPO therefore improved some dimensions of grounding without uniformly improving every type of reasoning claim.

No trace was fully grounded under the strict record-level criterion. However, all 277 GRPO outputs included an uncalibrated confidence claim and therefore necessarily received at least a minor issue. More importantly, 25.6% still contained a substantive or critical issue, and these traces exhibited poorer interval agreement. ROSCOE showed increased step entailment and chain faithfulness after GRPO, indicating stronger internal coherence but not guaranteed factual support. Coherence, grounding, trace–answer consistency, and predictive accuracy should therefore be evaluated as related but distinct properties.

Overall, reasoning-based training can improve selected biological prediction objectives even when the available evidence is incomplete, heterogeneous, and less directly verifiable than the premises used in formal mathematical or physical problems. The analogy is not that biological reasoning follows equally complete rules, but that intermediate reasoning and reinforcement learning can still improve structured prediction when paired with domain-specific representations and grounding checks. The strongest performance arose from combining language-model reasoning with a specialized residue classifier rather than requiring the generative model to solve every prediction objective alone.

Several limitations remain. Both cohorts contained only human IDR-positive proteins, so protein-level specificity and performance on fully ordered proteins were not measured. Teacher traces explained supplied answers rather than discovering IDRs de novo, and the underlying annotations remain incomplete. The zero-shot comparison evaluates the complete evidence-guided reasoning framework; a matched answer-only training condition would be needed to isolate the contribution of reasoning-trace supervision. Reward ablations used a single selected configuration, while GPT-6 Astra was an automated evaluator requiring validation by blinded experts. Future work should include multi-species and IDR-negative proteins, multiple training seeds, calibrated confidence values, direct boundary-justification objectives, and richer structural representations. Applying the framework to binding sites, functional regions, and protein interactions would further test whether its benefits extend beyond IDR prediction.

AI USE STATEMENT

Generative AI was used in two roles. First, as part of the research method, GPT-5.4 generated teacher reasoning traces; Qwen models were trained through supervised fine-tuning and GRPO to generate reasoning traces and IDR predictions; and GPT-6 Astra provided an automated evaluation of trace grounding. Their roles, inputs, outputs, and validation procedures are documented in the main text and appendix. Second, AI assistants provided support for code review and language editing. The authors reviewed all AI-assisted code and text, executed and verified the reported analyses, determined the scientific interpretations, and take responsibility for the final manuscript and its results.

ETHICS STATEMENT

This study uses publicly available protein-sequence and annotation resources and does not involve human participants, personally identifiable information, clinical interventions, animal experiments, or newly collected biological samples. The language models were accessed in accordance with their applicable licenses and terms of use. The framework is intended for computational research on intrinsically disordered protein regions, not for clinical or diagnostic use. Because generated explanations may contain unsupported claims, we report explicit grounding audits and emphasize that predictions require expert review and experimental validation before biological or translational use.

REPRODUCIBILITY STATEMENT

The main text specifies the data sources, cohort-construction and exclusion rules, model versions, SFT and GRPO hyperparameters, fusion equation, decoder settings, evaluation metrics, and perturbation procedures. The appendix provides detailed model comparisons, reward ablations, grounding results, and representative prompts and structured outputs. Cohort membership and accession- and sequence-level exclusions were fixed before independent evaluation, and neither locked evaluation cohort was used to tune the auxiliary head, fusion weight, or decoder. Because GPT-5.4 and GPT-6 Astra are hosted models, exact regeneration depends on continued access to the specified model versions; we therefore document their prompts, retained outputs, quality controls, and evaluation criteria.

REFERENCES

Adrian M. Altenhoff, Alex Warwick Vesztrocy, Charles Bernard, Clément-Marie Train, Alina Nicheperovich, Silvia Prieto Baños, Irene Julca, David Moi, Yannis Nevers, Sina Majidian, Christophe Dessimoz, and Natasha M. Glover. OMA orthology in 2024: Improved prokaryote coverage, ancestral and extant GO enrichment, a revamped synteny viewer and more in the OMA ecosystem. *Nucleic Acids Research*, 52(D1):D513–D521, 2024. doi: 10.1093/nar/gkad1020.

Maria Cristina Aspromonte, Maria Victoria Nugnes, Federica Quaglia, Adel Bouharoua, DisProt Consortium, Silvio C. E. Tosatto, and Damiano Piovesan. DisProt in 2024: Improving function annotation of intrinsically disordered proteins. *Nucleic Acids Research*, 52(D1):D434–D441, 2024. doi: 10.1093/nar/gkad928.

M. Madan Babu, Robin van der Lee, Natalia Sanchez de Groot, and Jörg Gsponer. Intrinsically disordered proteins: Regulation and disease. *Current Opinion in Structural Biology*, 21(3):432–440, 2011. doi: 10.1016/j.sbi.2011.03.011.

Salman F. Banani, Hyun O. Lee, Anthony A. Hyman, and Michael K. Rosen. Biomolecular condensates: Organizers of cellular biochemistry. *Nature Reviews Molecular Cell Biology*, 18(5): 285–298, 2017. doi: 10.1038/nrm.2017.7.

Yuhang Chen, Zhen Tan, Ruichen Zhang, Mufan Qiu, and Tianlong Chen. CellDuality: Unlocking biological reasoning in LLMs with self-supervised RLVR. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=I4meJN28Ol.

Adibvafa Fallahpour, Arman Seyed-Ahmadi, Parsa Idehpour, Omar Ibrahim, Benedict M. H. Choi, Purav Gupta, Jack Naimer, Kevin Zhu, Arnav Shah, Shihao Ma, Abhinav Adduri, Talu Güloglu, Nuo Liu, Haotian Cui, Arihant Jain, Max de Castro, Amirfaham Fallahpour, Antonio Cembellin-Prieto, John S. Stiles, Filip Nemčko, Alexander A. Nevue, Hyungseok C. Moon, Lucas Sosnick, Olivia Markham, Haonan Duan, Michelle Y. Y. Lee, Andrea F. M. Salvador, Chris J. Maddison, Christoph A. Thaiss, Chiara Ricci-Tam, Brian S. Plosky, Dave P. Burke, Patrick D. Hsu, Hani Goodarzi, and Bo Wang. BioReason-Pro: Advancing protein function prediction with multimodal biological reasoning. *bioRxiv*, 2026. doi: 10.64898/2026.03.19.712954. URL https://www.biorxiv.org/content/10.64898/2026.03.19.712954.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International conference on machine learning*, pp. 10835–10866. PMLR, 2023.

Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. Roscoe: A suite of metrics for scoring step-by-step reasoning. *arXiv preprint arXiv:2212.07919*, 2022.

Javier Herrero, Matthieu Muffato, Kathryn Beal, Stephen Fitzgerald, Leo Gordon, Miguel Pignatelli, Albert J. Vilella, Stephen M. J. Searle, Ridwan Amode, Simon Brent, William Spooner, Eugene Kulesha, Andrew Yates, and Paul Flicek. Ensembl comparative genomics resources. *Database*, 2016:bav096, 2016. doi: 10.1093/database/bav096.

A. Holehouse and B. Kragelund. The molecular basis for cellular function of intrinsically disordered regions. *Nature reviews. Molecular cell biology*, 25:187 – 211, 2023. doi: 10.1038/s41580-023-00673-0.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Mingyu Jin, Haochen Xue, Zhenting Wang, Boming Kang, Ruosong Ye, Kaixiong Zhou, Mengnan Du, and Yongfeng Zhang. ProLLM: Protein chain-of-thoughts enhanced LLM for protein-protein interaction prediction. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=2nTzomzjjb.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Zídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Md Wasi Ul Kabir and Md Tamjidul Hoque. Dispredict3.0: Prediction of intrinsically disordered regions/proteins using protein language model. *Applied Mathematics and Computation*, 472: 128630, 2024. doi: 10.1016/j.amc.2024.128630.

Takeshi Kojima, Shixiang Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1613. URL https://proceedings.neurips.cc/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html.

Benjamin J. Liebeskind, Claire D. McWhite, and Edward M. Marcotte. Towards consensus gene ages. *Genome Biology and Evolution*, 8(6):1812–1823, 2016. doi: 10.1093/gbe/evw113.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

Chaozhong Liu, Linlin Chao, Shaomin Ji, Hao Wang, Taorui Jiang, Zhangyang Gao, Yucheng Guo, Ming Yang, and Xiaoming Zhang. ProteinReasoner: A multi-modal protein language model with chain-of-thought reasoning for efficient protein design. *bioRxiv*, art. 2025.07.21.665832, 2025. doi: 10.1101/2025.07.21.665832. URL https://www.biorxiv.org/content/10.1101/2025.07.21.665832v3.

Mahta Mehdiabadi, Alessio Del Conte, Maria Victoria Nugnes, Maria Cristina Aspromonte, Silvio C. E. Tosatto, and Damiano Piovesan. Critical assessment of protein intrinsic disorder round 3: Predicting disorder in the era of protein language models. *Proteins: Structure, Function, and Bioinformatics*, 94(1):414–424, 2026. doi: 10.1002/prot.70045.

Marco Necci, Damiano Piovesan, CAID Predictors, DisProt Curators, and Silvio C E Tosatto. Critical assessment of protein intrinsic disorder prediction. *Nat. Methods*, 18(5):472–481, May 2021. doi: 10.1038/s41592-021-01117-3.

Damiano Piovesan, Alessio Del Conte, Mahta Mehdiabadi, Maria Cristina Aspromonte, Matthias Blum, Giulio Tesei, Sören von Bülow, Kresten Lindorff-Larsen, and Silvio C. E. Tosatto. MOBIDB in 2025: Integrating ensemble properties and function annotations for intrinsically disordered proteins. *Nucleic Acids Research*, 53(D1):D495–D503, 2025. doi: 10.1093/nar/gkae969.

Kiersten M. Ruff and R. Pappu. Alphafold and implications for intrinsically disordered proteins. *Journal of molecular biology*, pp. 167208, 2021. doi: 10.1016/j.jmb.2021.167208.

Marc D. Singleton and M. Eisen. Evolutionary analyses of intrinsically disordered regions reveal widespread signals of conservation. *PLOS Computational Biology*, 20:e1012028 – e1012028, 2024. doi: 10.1371/journal.pcbi.1012028.

G. Tesei, Anna Ida Trolle, Nicolas Jonsson, Johannes Betz, Frederik E Knudsen, Francesco Pesce, K. E. Johansson, and K. Lindorff-Larsen. Conformational ensembles of the human intrinsically disordered proteome. *Nature*, 626:897 – 904, 2024. doi: 10.1038/s41586-023-07004-5.

The UniProt Consortium. UniProt: The universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 2023. doi: 10.1093/nar/gkac1052.

Mihaly Varadi et al. AlphaFold Protein Structure Database in 2024: Providing structure coverage for over 214 million protein sequences. *Nucleic Acids Research*, 52(D1):D368–D375, 2024. doi: 10.1093/nar/gkad1011.

Kui Wang, Gang Hu, Sushmita Basu, and Lukasz Kurgan. fldpnn2: Accurate and fast predictor of intrinsic disorder in proteins. *Journal of Molecular Biology*, 436(17):168605, 2024. doi: 10.1016/j.jmb.2024.168605.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Peter E. Wright and H. Jane Dyson. Intrinsically disordered proteins in cellular signalling and regulation. *Nature Reviews Molecular Cell Biology*, 16(1):18–29, 2015. doi: 10.1038/nrm3920.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report, 2024. URL https://arxiv.org/abs/2412.15115.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

APPENDIX

A ADDITIONAL PERFORMANCE RESULTS

Table A.1: Performance of Qwen3-4B hybrid and baseline models on the locked teacher-corpus validation and independent test cohorts ($n = 277$ each).

Cohort	Model	AUROC	AUPRC	$F_{\max}$	MCC@$F_{\max}$	Residue F1	Residue MCC	Interval F1	Best IoU
Teacher-corpus validation									
	SFT hybrid, $\alpha = 0.5$	0.9840	0.9848	0.9455	0.8873	0.9474	0.8903	0.8759	0.7830
	GRPO hybrid, $\alpha = 0.5$	0.9842	0.9851	0.9451	0.8871	0.9463	0.8881	0.8718	0.7795
	flDPnn2	0.9290	0.9268	0.8656	0.7123	0.6950	0.5572	–	–
	DisPredict3.0	0.8599	0.8742	0.8073	0.5935	0.4713	0.3911	–	–
	Zero-shot Qwen, sequence only	–	–	–	–	0	–	0	0
	Zero-shot Qwen, sequence + evidence	–	–	–	–	0	–	0	0
Independent test									
	SFT hybrid, $\alpha = 0.5$	0.9323	0.8986	0.8387	0.7378	0.8423	0.7439	0.7399	0.5224
	GRPO hybrid, $\alpha = 0.5$	0.9305	0.8928	0.8354	0.7320	0.8377	0.7336	0.7480	0.5495
	flDPnn2	0.8670	0.7983	0.7535	0.5920	0.5094	0.4384	–	–
	DisPredict3.0	0.8465	0.7635	0.7324	0.5500	0.2547	0.2678	–	–

AUROC, AUPRC, $F_{\max}$, and MCC at $F_{\max}$ were calculated from pooled residue scores. Hybrid residue F1 and MCC used the frozen normalized decoder; flDPnn2 and DisPredict3.0 used their returned binary labels. Interval F1 and mean best IoU were not evaluated for the external residue-level predictors. Zero-shot Qwen was evaluated only on the teacher-corpus validation set and did not produce usable intervals. The teacher-corpus validation set was candidate-enriched, whereas independent-test selection did not use candidate/reference agreement.

B ABLATION STUDIES

Hyperparameter search Among the configurations considered, group size 8 with learning rate 2.5×10^{-5} achieves the highest residue-level F1 at 0.7968 (Table B.1) and a reasonably high mean IoU. At the same learning rate, group size 16 yields lower F1 and IoU than group size 8, supporting the selection of group size 8. The 1,690-step checkpoint is confirmed using the two-part stopping criterion described above. Compared to step 200, training to step 1,690 yields substantial improvements in both F1 ($\Delta = +0.0725$, $p = 3.1 \times 10^{-6}$) and IoU ($\Delta = +0.0885$, $p = 4.2 \times 10^{-10}$). Relative to step 1,650, the step-1,690 checkpoint shows only marginal gains ($\Delta\text{F1} = 0.0023$, $\Delta\text{IoU} = 0.0009$), indicating that evaluation performance had plateaued. Together, these criteria define group size 8, learning rate 2.5×10^{-5}, and the 1,690-step checkpoint as the reference configuration.

Table B.1: Raw generative performance after 1,690 training steps across learning rates and group sizes. F1 is pooled residue-level F1 and IoU is mean best interval IoU.

Learning rate	5×10^{-6}		2.5×10^{-5}		1×10^{-4}	
Group size	Residue F1	IoU	Residue F1	IoU	Residue F1	IoU
SFT baseline	F1 = 0.6465, IoU = 0.6245					
8	0.7476	0.6311	**0.7968**	**0.7313**	0.7761	0.7366
16	0.7495	0.6772	0.7874	0.7131	0.7631	0.6893

Agreement with source-specific disorder evidence The selected reinforcement learning model, trained for 1,690 steps with a group size of 8 and a learning rate of 2.5×10^{-5}, agrees more closely with several source-specific annotations and proxies than the SFT checkpoint from which it was initialized (Table B.2). These comparisons include DisProt, MobiDB-lite, and a low-pLDDT flexibility proxy. They are not fully independent benchmarks because DisProt and MobiDB contributed to construction of the reference corpus; the analysis instead tests source-specific agreement under a common evaluation procedure. The teacher traces were generated by GPT-5.4 from the visible protein evidence and the supplied outcome label. Residue-level Jaccard increases across all three sources, with the largest gains for DisProt and the low-pLDDT proxy. Reference coverage increases

for DisProt but decreases slightly for MobiDB-lite and the low-pLDDT proxy, consistent with a more conservative predictor. SFT estimates use only examples with a parsable answer interval, whereas teacher and GRPO traces provide intervals for all source-covered examples; comparisons across systems should therefore be interpreted cautiously. The DisProt estimate is also based on only $n = 20$ source-covered examples.

The ROSCOE evaluation assesses reasoning-trace quality along complementary dimensions. On the entailment-based measures, the selected RL checkpoint performs substantially better than both the outcome-informed teacher traces and the SFT baseline. Its step entailment reaches 0.1574, about twice the teacher value, while chain faithfulness increases to 0.3415 compared with 0.0520 for the teacher traces. The embedding-based measures show a different pattern. Step similarity is slightly lower for RL, while answer coverage is nearly unchanged across the three systems. These results indicate that RL improves some aspects of trace quality, particularly entailment and consistency across reasoning steps, without compromising embedding-based similarity or coverage.

Table B.2: Agreement with source-specific disorder annotations and proxies.

(a) Mean residue-level Jaccard with evaluable sample counts in parentheses.

Evidence source (source-covered n)	Teacher	SFT	Selected RL checkpoint
DisProt (20)	0.968 (20)	0.295 (16)	0.453 (20)
MobiDB-lite (189)	0.540 (189)	0.429 (142)	0.467 (189)
Low-pLDDT flexibility proxy (270)	0.800 (270)	0.558 (210)	0.635 (270)

(b) Mean reference-residue coverage.

Evidence source (source-covered n)	Teacher	SFT	Selected RL checkpoint
DisProt (20)	0.969	0.340	0.488
MobiDB-lite (189)	0.541	0.516	0.493
Low-pLDDT flexibility proxy (270)	0.814	0.738	0.701

(c) ROSCOE evaluation of reasoning traces.

ROSCOE metric	Teacher trace	SFT baseline	Selected RL checkpoint
Step entailment	0.0775	0.0656	0.1574
Chain faithfulness	0.0520	0.0193	0.3415
Step similarity	0.1954	0.1933	0.1569
Answer coverage	0.2471	0.2333	0.2349

Reward specification. Holding the architecture, data, and training budget fixed, we explored how reward specification influences prediction performance. The experimental design is shown in Table B.3. Three conditions reweighted the existing answer, format, and consistency terms, while three others added a *source-agreement* term based on DisProt, MobiDB, and the low-pLDDT proxy, assigning a neutral score when no comparison was available. Each condition trained a separate policy from the same SFT checkpoint using one random seed; the comparisons are therefore descriptive.

Source-agreement evidence was available for 271 of 277 development proteins (97.8%). Incorporating it into the reward produced modest, nonuniform gains, with different weights benefiting different outcomes (Table B.4). Overall, the ablation provides limited support for replacing the default reward specification.

Table B.3: Reward weights used for reward specification ablation experiments

Reward weights	Default	Format+	Answer+	Source(.1)	Source(.3)	Source(.6)
Answer	1.0	1.0	1.0	1.0	1.0	1.0
Format	0.2	**0.5**	**0.1**	0.2	0.2	0.2
Consistency	0.1	0.1	**0.05**	0.1	0.1	0.1
Source agreement	0	0	0	**0.1**	**0.3**	**0.6**

Reward combines answer, format, consistency, and source-agreement terms; the answer score combines IoU, F1, and interval-count match. Bold indicates values changed from the default configuration. The pre-RL SFT baseline is omitted.

Table B.4: IDR prediction performance in reward specification ablation experiments

Condition	F1	IoU	Int.F1	JSON	AUROC	AUPRC	mIDR
SFT	0.6465	0.6245	0.7246	0.6318	0.5938	0.4857	0.2191
default	0.8458	0.7313	**0.7968**	**1.0000**	**0.8732**	**0.8531**	**0.7725**
format+	0.8311	0.7316	0.7795	**1.0000**	0.8594	0.8418	0.7503
answer+	0.8402	0.7295	0.7729	**1.0000**	0.8349	0.7749	0.7630
ext-ev (0.1)	**0.8482**	**0.7434**	0.7880	**1.0000**	0.8535	0.8024	0.7565
ext-ev (0.3)	0.8445	0.7360	0.7806	**1.0000**	0.8388	0.7928	0.7472
ext-ev (0.6)	0.8382	0.6893	0.7631	**1.0000**	0.8522	0.8148	0.6975

F1 = residue F1 (macro), IoU = mean best-match IoU, Int.F1 = interval F1, JSON = valid-JSON rate, mIDR = multi-IDR F1. The residue F1 reported here is macro-averaged across proteins, whereas Table B.1 uses pooled residue F1; the two values are not directly comparable. format+ and answer+ denote the format-heavy and answer-sharpened variants; ext-ev(w) denotes the source-agreement term at weight w.

Table B.5: GPT-6 Astra evaluation of evidence grounding in SFT- and GRPO-generated Qwen reasoning traces and answers for the same 277-protein validation cohort. Prediction decisions, boundary-justification gaps, confidence calibration, unsupported biological rationales, and fabricated or contradicted evidence are evaluated separately.

Metric	SFT	GRPO
Protein-level grounding judgment		
Fully grounded	0 (0.00%)	0 (0.00%)
Minor issue	193 (69.68%)	206 (74.37%)
Substantive grounding issue	74 (26.71%)	64 (23.10%)
Fabricated or contradicted evidence	10 (3.61%)	7 (2.53%)
Substantive or critical issue	84 (30.32%)	71 (25.63%)
Trace–answer consistent	193 (69.68%)	250 (90.25%)
Valid generated JSON	277 (100.00%)	277 (100.00%)
Claim-level classification		
Classified claims	1,965	1,920
Supported statement	608 (30.94%)	548 (28.54%)
Reasonable inference	330 (16.79%)	336 (17.50%)
Prediction decision	337 (17.15%)	334 (17.40%)
Confidence-calibration issue	277 (14.10%)	277 (14.43%)
Boundary-justification gap	233 (11.86%)	240 (12.50%)
Unsupported biological rationale	126 (6.41%)	173 (9.01%)
Fabricated evidence	7 (0.36%)	8 (0.42%)
Contradicted evidence	15 (0.76%)	3 (0.16%)
Unverifiable	31 (1.58%)	1 (0.05%)
Not factual	1 (0.05%)	0 (0.00%)
Claim severity		
No issue	1,283 (65.29%)	1,219 (63.49%)
Minor	581 (29.57%)	617 (32.14%)
Substantive	89 (4.53%)	76 (3.96%)
Critical	12 (0.61%)	8 (0.42%)
Overall interval agreement		
Exact interval match	66 (23.83%)	60 (21.66%)
IoU ≥ 0.50	180 (64.98%)	195 (70.40%)
IoU ≥ 0.75	125 (45.13%)	133 (48.01%)
Mean best IoU	0.6399	0.6749
Median best IoU	0.6757	0.7222
Interval agreement by grounding label		
Minor issue: proteins	193	206
Minor issue: mean best IoU	0.7246	0.7340
Minor issue: exact match	60 (31.09%)	59 (28.64%)
Substantive issue: proteins	74	64
Substantive issue: mean best IoU	0.4531	0.5143
Substantive issue: exact match	6 (8.11%)	1 (1.56%)
Fabricated/contradicted: proteins	10	7
Fabricated/contradicted: mean best IoU	0.3867	0.4029
Fabricated/contradicted: exact match	0 (0.00%)	0 (0.00%)

C SUPPLEMENTARY TEXT

C.1 GPT-5.4 TEACHER-TRACE GENERATION

GPT-5.4 generated target-conditioned explanations for supervised training. It received the visible protein evidence together with the reference intervals and confidence values and was instructed to explain the supplied answer without referring to hidden annotation sources. Each output contained six reasoning operations: sequence-biophysics assessment, candidate-zone boundary assessment, structural-conflict assessment, AlphaFold pLDDT assessment, evolutionary assessment, and final boundary decision.

To reduce token usage, the compact teacher-corpus prompt omitted the literal amino-acid sequence while retaining sequence-derived biophysical descriptions, candidate zones, structural counterevidence, and available evolutionary evidence. This stage therefore generated explanations of supplied targets rather than independent predictions.

Visible protein evidence for A6NGK3.

```
Protein accession: A6NGK3
Gene: GAGE10
Protein name: G antigen 10 (GAGE-10)
Species: Homo sapiens (Human)
Length: 116 amino acids

Protein sequence: [provided in original prompt]

Biophysical and chemical sequence descriptions:
- residues 101-115: mean hydropathy -2.47, net charge/residue -0.13, charged fraction 0.53,
↪  polar fraction 0.20, aromatic fraction 0.00, Pro/Gly/Ser/Gln fraction 0.40,
↪  low-complexity score 0.18, sequence-only IDR tendency score 0.92
- residues 108-116: mean hydropathy -2.23, net charge/residue -0.22, charged fraction 0.44,
↪  polar fraction 0.44, aromatic fraction 0.00, Pro/Gly/Ser/Gln fraction 0.44,
↪  low-complexity score 0.21, sequence-only IDR tendency score 0.92
- residues 41-49: mean hydropathy -2.16, net charge/residue -0.22, charged fraction 0.44,
↪  polar fraction 0.22, aromatic fraction 0.00, Pro/Gly/Ser/Gln fraction 0.33,
↪  low-complexity score 0.19, sequence-only IDR tendency score 0.90
- residues 102-116: mean hydropathy -2.19, net charge/residue -0.13, charged fraction 0.53,
↪  polar fraction 0.27, aromatic fraction 0.00, Pro/Gly/Ser/Gln fraction 0.33,
↪  low-complexity score 0.17, sequence-only IDR tendency score 0.89
- residues 41-61: mean hydropathy -1.93, net charge/residue -0.33, charged fraction 0.43,
↪  polar fraction 0.19, aromatic fraction 0.00, Pro/Gly/Ser/Gln fraction 0.33,
↪  low-complexity score 0.17, sequence-only IDR tendency score 0.87

Merged sequence-derived candidate IDR zones:
- 1-116 (mean score 0.784, hydropathy -1.442, charged fraction 0.346, aromatic fraction
↪  0.026, Pro/Gly/Ser/Gln fraction 0.397)

Structure/domain conflict descriptions:
- No UniProt domain, transmembrane, or signal-peptide overlap is currently detected for the
↪  visible candidate zones.

Evolutionary descriptions:
- Whole-protein/gene age clade: Mammalia; approximate age: 225 MYA; age-call entropy: -0.0;
↪  node error: 0.866666666667.
- Region-level ortholog/MSA features are not yet available for the visible candidate zones.

You are a protein expert. Given the protein sequence and descriptions, does the sequence
↪  contain intrinsically disordered regions (IDRs) and if so, at which residues?
```

GPT-5.4 teacher trace and answer.

```
{
  "reasoning_trace": [
    {
      "step": 1,
      "operation": "sequence_biophysics_assessment",
      "observation": "The sequence is globally hydrophilic, charged, low in aromatics, and
      ↪  enriched in disorder-promoting residues.",
      "effect_on_idr_call": "supports"
    },
    {
      "step": 2,
      "operation": "candidate_zone_boundary_assessment",
```

```
      "observation": "Sequence analysis merges disorder tendency across the full 1-116
      ↪  length into one continuous candidate zone.",
      "effect_on_idr_call": "supports"
    },
    {
      "step": 3,
      "operation": "structure_conflict_assessment",
      "observation": "No annotated domain, transmembrane segment, or signal peptide
      ↪  conflicts with the full-length candidate region.",
      "effect_on_idr_call": "supports"
    },
    {
      "step": 4,
      "operation": "alphafold_plddt_assessment",
      "observation": "No pLDDT information is visible here. pLDDT is supporting evidence,
      ↪  not proof of disorder.",
      "effect_on_idr_call": "neutral"
    },
    {
      "step": 5,
      "operation": "evolutionary_assessment",
      "observation": "Gene age is Mammalia, but no region-level conservation features are
      ↪  available for this candidate zone.",
      "effect_on_idr_call": "neutral"
    },
    {
      "step": 6,
      "operation": "final_boundary_decision",
      "observation": "Visible evidence supports one full-length IDR spanning residues 1-116
      ↪  with high confidence.",
      "effect_on_idr_call": "supports"
    }
  ],
  "answer": {
    "contains_idr": true,
    "idr_intervals": [
      {
        "start": 1,
        "end": 116,
        "confidence": 0.856
      }
    ],
    "answer_text": "Answer: Yes. IDR residue interval(s): 1-116 (confidence 0.856). The
    ↪  visible sequence-derived candidate zone(s) are: 1-116 (mean score 0.784, hydropathy
    ↪  -1.442, charged fraction 0.346, aromatic fraction 0.026, Pro/Gly/Ser/Gln fraction
    ↪  0.397)"
  }
}
```


C.2 QWEN3 GRPO INFERENCE

At inference, the GRPO policy received the biological evidence without the reference answer, reference intervals, or teacher trace. Qwen additionally received up to eight projected ESM-2 candidate-region soft tokens. It generated five reasoning operations: sequence-biophysics assessment, candidate-zone and boundary assessment, structural-conflict assessment, combined AlphaFold and evolutionary assessment, and final boundary decision.

Target-blind inference prompt for A6NGK3. The following prompt was presented to Qwen without the reference answer, reference intervals, or GPT-5.4 teacher trace. The bracketed sequence phrase below is the literal placeholder retained in the archived compact prompt; sequence information was supplied separately through the ESM-2 soft tokens and the displayed sequence-derived features.

```
SYSTEM:
You are a protein expert. Given a protein sequence and feature
descriptions, produce a structured IDR reasoning trace and final
answer. Use only visible evidence in the prompt. Treat AlphaFold pLDDT
cautiously: low pLDDT is supporting evidence, not proof of disorder;
high pLDDT is counterevidence but does not fully exclude conditional
disorder.

USER:
Protein accession: A6NGK3
Gene: GAGE10
Protein name: G antigen 10 (GAGE-10)
```

```
Species: Homo sapiens (Human)
Length: 116 amino acids

Protein sequence: [provided in original prompt]

Biophysical and chemical sequence descriptions:
- residues 101-115: mean hydropathy -2.47, net charge/residue -0.13,
  charged fraction 0.53, polar fraction 0.20, aromatic fraction 0.00,
  Pro/Gly/Ser/Gln fraction 0.40, low-complexity score 0.18,
  sequence-only IDR tendency score 0.92
- residues 108-116: mean hydropathy -2.23, net charge/residue -0.22,
  charged fraction 0.44, polar fraction 0.44, aromatic fraction 0.00,
  Pro/Gly/Ser/Gln fraction 0.44, low-complexity score 0.21,
  sequence-only IDR tendency score 0.92
- residues 41-49: mean hydropathy -2.16, net charge/residue -0.22,
  charged fraction 0.44, polar fraction 0.22, aromatic fraction 0.00,
  Pro/Gly/Ser/Gln fraction 0.33, low-complexity score 0.19,
  sequence-only IDR tendency score 0.90
- residues 102-116: mean hydropathy -2.19, net charge/residue -0.13,
  charged fraction 0.53, polar fraction 0.27, aromatic fraction 0.00,
  Pro/Gly/Ser/Gln fraction 0.33, low-complexity score 0.17,
  sequence-only IDR tendency score 0.89
- residues 41-61: mean hydropathy -1.93, net charge/residue -0.33,
  charged fraction 0.43, polar fraction 0.19, aromatic fraction 0.00,
  Pro/Gly/Ser/Gln fraction 0.33, low-complexity score 0.17,
  sequence-only IDR tendency score 0.87

Merged sequence-derived candidate IDR zones:
- 1-116 (mean score 0.784, hydropathy -1.442,
  charged fraction 0.346, aromatic fraction 0.026,
  Pro/Gly/Ser/Gln fraction 0.397)

Structure/domain conflict descriptions:
- No UniProt domain, transmembrane, or signal-peptide overlap is
  currently detected for the visible candidate zone.

Evolutionary descriptions:
- Whole-protein/gene age clade: Mammalia; approximate age: 225 MYA;
  age-call entropy: -0.0; node error: 0.866666666667.
- Region-level ortholog/MSA features are not yet available for the
  visible candidate zone.

Given the protein sequence and descriptions, does the sequence contain
intrinsically disordered regions (IDRs), and if so, at which residues?

Instruction:
In the final answer, report both the number of IDRs and their residue
intervals on the input protein sequence.

Copy Protein ID and Protein accession exactly into protein_id and
uniprot_accession. Do not infer, substitute, or memorize different IDs.
Return one JSON object containing a five-step reasoning_trace followed
by answer.
```

GRPO reasoning trace and raw answer.


```
{
  "reasoning_trace": [
    {
      "step": 1,
      "operation": "sequence_biophysics_assessment",
      "observation": "Sequence is polar and low hydropathy."
    },
    {
      "step": 2,
      "operation": "candidate_zone_boundary_assessment",
      "observation": "Target boundaries 1-116; retain 1-116."
    },
    {
      "step": 3,
      "operation": "structure_conflict_assessment",
      "observation": "No annotated domains, transmembranes, or signal peptides overlap."
    },
    {
      "step": 4,
      "operation": "alphafold_and_evolutionary_assessment",
      "observation": "No pLDDT data are shown here."
    },
    {
      "step": 5,
```

```
      "operation": "final_boundary_decision",
      "observation": "Final"
    }
  ],
  "answer": {
    "contains_idr": true,
    "num_idrs": 1,
    "idr_intervals": [
      {
        "start": 1,
        "end": 116,
        "confidence": 1
      }
    ],
    "answer_text": "Answer: Yes."
  }
}
```


C.3 Matched Hybrid Prediction Example

A6NGK3 was selected post hoc from the teacher-corpus validation set as a clear illustrative exact-match example and should not be interpreted as representative of average model performance. The raw GRPO answer and final hybrid prediction both identified one full-length IDR spanning residues 1–116, exactly matching the reference interval.

The LLM-as-judge assigned the trace a minor grounding issue. Its biological claims were supported by the visible evidence, while the numerical confidence was flagged as unsupported because it was not produced by an explicitly calibrated confidence model.

For the final hybrid prediction, the GRPO interval proposal was fused with frozen auxiliary-head probabilities using $\alpha = 0.5$. The prespecified decoder used a threshold of 0.4647, smoothing width of one residue, minimum interval length of five residues, and maximum merge gap of five residues. The final interval was unchanged from the raw GRPO proposal, and postprocessing did not modify the reasoning trace.

Final hybrid answer and reference interval.


```
{
  "final_hybrid_answer": {
    "contains_idr": true,
    "num_idrs": 1,
    "idr_intervals": [
      {
        "start": 1,
        "end": 116,
        "confidence": 0.992284
      }
    ],
    "answer_text": "Answer: Yes. Number of IDRs: 1. IDR residue interval(s): [1-116]."
  },
  "reference_intervals": [
    {
      "start": 1,
      "end": 116
    }
  ],
  "postprocessing": {
    "method": "frozen_auxiliary_head_grpo_interval_fusion_and_decoder",
    "fusion_alpha": 0.5,
    "decoder_parameters": {
      "threshold": 0.4647,
      "smoothing_width": 1,
      "min_length": 5,
      "merge_gap": 5
    },
    "trace_modified": false,
    "raw_parse_status": "object",
    "raw_identity_status": "matched",
    "raw_vs_final_interval_relation": "unchanged",
    "trace_answer_consistency_review_required": false,
    "grpo_interval_status": "object"
  }
}
```