

Enhancing protein structure prediction accuracy by prioritizing important residues using protein language models

Corresponding Author: Dr Yu Liu

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Comments to the Author

This manuscript presents i-Fold, a modified AlphaFold2-based framework that incorporates residue importance scores (RIS) derived from ESM-2 into the training loss. The idea of guiding structure prediction using residue-level importance is interesting, and the authors provide benchmark and independent dataset evaluations showing improved performance compared with AlphaFold2. Overall, the manuscript is clearly written, and the study is potentially useful to researchers working on protein structure prediction.

I have several suggestions that may help improve clarity, presentation, and the completeness of the evaluation.

1. Fig. 1 would benefit from a clearer illustration of the overall pipeline. It is currently, RIS are generated, where they are introduced into the model, and which components they influence during prediction. A revised schematic that explicitly shows the RIS module and its points of interaction with the AlphaFold2 pipeline would improve readability and make the method easier to follow.
2. The manuscript evaluates prediction accuracy mainly using RMSD. While RMSD is informative, it is sensitive to local structural deviations and protein length. Standard structure prediction studies typically report additional metrics such as TM-score or IDDT. Including these metrics would provide a more comprehensive assessment.
3. Currently, the method is compared only with AlphaFold2. If possible, additional comparisons with more recent structure prediction methods, such as AlphaFold3, Chai-1, Boltz-1, etc., would better position the contribution of i-Fold and strengthen the conclusions.
4. The training procedure is described only superficially. Key details such as batch size, number of steps or epochs, recycling settings, and number of trained model parameter sets are not provided. The description of training duration also appears incomplete. Dataset preparation is similarly unclear: although the authors state that 185,158 structures were used for training and 3,599 sequences for testing, it is not specified whether redundancy filtering or homology-based separation was applied. Without this information, potential data leakage cannot be ruled out.
5. When comparing iFold and AlphaFold2 modeling performance, please clarify whether the number of generated samples (or recycling/sampling strategies) is controlled to be consistent across methods. If not already stated, the authors should explicitly describe the sampling protocol to ensure the fairness of the comparisons.
6. Additional ablation experiments would substantially strengthen the methodological claims. In particular:
 - (1) A control experiment using randomized RIS weights (if computationally feasible) to quantify the real contribution of RIS.
 - (2) A clear comparison of different training strategies, e.g., structure-module-only fine-tuning vs. full-model retraining, if the current model involves full retraining.
7. The explanation for the poor performance on target 7JW6 relies on the concept of "evolutionary noise" generated by ESM. As this is a singular anecdotal example, the manuscript would benefit from a quantitative analysis determining the frequency of such RIS misweighting events. Furthermore, the authors should investigate if there is a systematic correlation between performance degradation and anomalous entropy distributions. Furthermore, the discussion could be strengthened by analyzing the underlying causes of these failures and proposing concrete improvement strategies, possibly accompanied by preliminary validation experiments.
8. There is a conflicting report regarding the benchmark dataset size. The Abstract cites 3,559 structures, whereas both the Results and Methods sections refer to 3,599 sequences. Please ensure these figures are consistent throughout the text.

9. In Fig. 2C, the axis labels ("Training epoch" and "Validation Loss") follow inconsistent capitalization conventions. The authors may consider standardizing figure label formatting and checking other figures for consistency.
10. Some quantitative axes appear to lack explicit units (e.g., Fig. 3C). For measurements such as structural deviation, indicating the unit would help readers interpret the scale of the results.
11. Several figures are visually coarse and do not clearly convey the methodology. For example, the overview in Fig. 1 does not clearly illustrate the computational pipeline or where RIS weighting is applied. Improving figure clarity would enhance the readability of the manuscript.
12. Several spelling errors throughout the manuscript and supplementary files should be addressed to improve professional clarity. Specifically, please correct "optimized" to "optimization" and "weighting" to "weighting" in the Methods section (Page 10), and fix "Funtion" to "Function" in the title of Supplementary File S5.

Reviewer #2

(Remarks to the Author)

The authors present i-Fold, an adaptation of the AlphaFold2 architecture that integrates residue-specific importance scores (RIS) derived from the ESM2 protein language model into the FAPE loss function. The manuscript's primary strength lies in its "simple yet effective" approach—achieving performance gains without adding new parameters or significantly increasing computational overhead. However, several critical methodological ambiguities regarding the training process, data controls, and the nature of the RIS metric must be addressed.

Major Revisions

1. In the section Methods-Model training, the authors state that models were "trained on four NVIDIA A100 GPUs... completing 10 over approximately seven days".

Is that 10 epochs? The author must clarify this sentence.

Original AlphaFold2 training required massive TPU resources over several weeks. Did the authors initialize the model with weights from a pre-trained OpenFold/AlphaFold2 checkpoint? If this is a fine-tuning exercise rather than a "retraining" from scratch as implied, the manuscript must be updated to reflect this, as the 10-epoch gain might represent localized optimization rather than a fundamental architectural shift.

2. The manuscript does not explicitly state whether structural templates were used during the training and prediction phase for i-Fold and the AlphaFold2 baseline. If templates were used, the observed improvements might stem from template-driven modeling rather than the RIS-weighted loss.

Please clarify whether structural templates were utilized in this work. Additionally, which specific version of AlphaFold2 (e.g., v2.0 vs v2.3) was used for comparison, and were the same MSA/template inputs provided to both models to ensure a fair comparison? Furthermore, please specify the number of models generated per target for both i-Fold and AlphaFold2. Was the model selected for evaluation the 'best' among these predictions, and if so, what were the specific criteria (e.g., RMSD) used for this selection?

3. In section 'RIS improves protein structure prediction through multiple mechanisms', the authors argue that RIS captures information beyond evolutionary conservation, citing the protein HGSNAT (8VKJ) protein as an example where AlphaFold2 failed despite being a "high-homology" sequence.

Do the authors use "raw hits" to reflect MSA diversity? Raw hits can be highly redundant and do not accurately reflect the Neff (Number of effective sequences). A target with 10,000 hits might still have a very low Neff, leading to AF2 failure.

Please provide the specific Neff values for the 8VKJ target. Furthermore, provide a formal definition of the "MSA-derived conservation score" used in the correlation analysis in Figure 5B to ensure the 0.45 Spearman's Rho is a meaningful baseline.

4. AlphaFold2 produces a pLDDT score (Predicted Local Distance Difference Test) as a measure of local confidence. Since RIS is intended to highlight functionally and structurally critical residues, there should be a logical relationship between these metrics.

Please provide a correlation analysis between the ESM-derived RIS and the model's final predicted pLDDT scores. If RIS is merely highlighting regions the model already identifies as high-confidence, its utility is redundant; if they are uncorrelated, the authors should explain why forcing the model to focus on "important" but "low-confidence" regions improves overall global accuracy.

Minor Points

In Figure 3, subfigures A, C, D, and E lack axis titles, which makes it difficult to interpret the results.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The authors have carefully and adequately addressed all of my previous comments and concerns. Their responses are clear, comprehensive, and satisfactory, and the revised manuscript has been improved accordingly. I have no further questions or suggestions. Therefore, I recommend that the manuscript be accepted for publication.

Reviewer #2

(Remarks to the Author)

1. The author acknowledged that the benchmark test set ($n = 3,599$) was randomly sampled without homology-based deduplication, and I agree that the relative improvement (RMSD $1.794 \rightarrow 1.482$ Å) remains informative as a matched-condition comparison.

However, I respectfully disagree with the claim that the independent test set of 167 newly released structures "completely eliminates the risk of data leakage." Release date alone is not a sufficient criterion: newly deposited PDB structures predominantly belong to previously characterized protein families, and their sequences typically have multiple homologs already present in the AlphaFold2 training data and in BFD/UniRef. The standard for a leakage-controlled test set is sequence identity to training data (commonly $< 30\%$ pairwise identity, or cluster-based splitting), not deposition date.

The authors already have the analytical tools required to address this. In the Methods, they define an "effective MSA hits" metric based on 0.8-similarity clustering of a3m alignments, and they apply this metric to the benchmark test set (Fig. 5, Supplementary Fig. S5). However, no equivalent analysis is reported for the independent test set of 167 structures. Reporting the effective MSA hits for each of the 167 independent test sequences would help clarify this.

2. Assuming the reported test results are obtained on samples held out independently from training, I largely agree with the authors' conclusion that i-Fold's significantly lower prediction error than the retrained AlphaFold2 baseline within the 10-epoch window supports the value of the RIS-weighted loss function.

That said, I would recommend that the authors add an estimate of the best performance i-Fold could be expected to achieve, based on the original OpenFold paper (Ahdriz et al., 2024, Nature Methods 21:1514–1524, doi: 10.1038/s41592-024-02272-z). OpenFold reports that validation IDDT-C α reaches ~ 0.83 within a few thousand training steps at an effective batch size of 132 per step, whereas i-Fold was trained for 12,500 steps (10 epochs $\times$ 1,250 steps, batch size = 8). OpenFold's "few thousand steps" (e.g., 5,000 steps $\times$ 132 = 660,000 samples) corresponds to approximately 40–90 epochs of i-Fold training under the current configuration. Including this reference point would help readers benchmark i-Fold's performance against the expected training trajectory.

3. I appreciate the authors' transparency regarding the relatively low correlation between RIS and pLDDT (median Spearman correlation coefficient: 0.29). However, I believe the response would be substantially strengthened by addressing two specific points that more directly engage with my original question.

First, the reported correlation ($\rho = 0.29$) is computed across all residues, but it would be informative to know whether this correlation differs between the two structural categories the authors themselves analyze. Specifically, what is the RIS–pLDDT correlation within secondary-structure regions versus within non-secondary-structure (flexible linker) regions? A meaningful difference between these two strata would clarify whether RIS provides distinct information primarily in one region type versus the other, and would lend further mechanistic support to the proposed "cascading effect."

Second, the core question — why forcing the model to focus on important but low-confidence regions improves overall global accuracy — is not directly addressed by the per-region RMSD comparison ($2.67 \rightarrow 2.13$ Å for secondary-structure regions, $2.80 \rightarrow 2.48$ Å for non-secondary-structure regions). To directly support the authors' claim that RIS improves global accuracy beyond regions of secondary structure, I recommend adding a concrete example: identify one or more residues (or contiguous regions) in the independent test set that have high RIS but low AlphaFold2 pLDDT, and demonstrate that i-Fold improves the local prediction accuracy at these specific positions. Such a case study — ideally with structural visualization — would provide the mechanistic intuition that a single correlation coefficient cannot convey, and would directly answer the reviewer's question of why redirecting attention to "important but uncertain" residues yields global rather than merely local improvements.

Version 2:

Reviewer comments:

Reviewer #2

(Remarks to the Author)

The authors have addressed most of my comments. I only have one minor suggestion regarding terminology.

I suggest avoiding the term "low-homology group" for sequences with effective MSA hits < 1000 , as this threshold does not necessarily indicate low homology in the conventional sense. A more precise description would be a medium-homology group, which would more accurately reflect what was actually measured.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

April 24th, 2026

Re: Manuscript No. COMMSBIO-26-0247

Dear Dr. Yang,

Thank you for giving us the opportunity to revise our paper. We would like to express our thanks to the reviewers for the constructive and positive comments. We have made revisions and additions to the manuscript format as per the following requirements, and we hope it meets the publishing standards. If there are any problems during the editing process, please feel free to contact me at any time.

We declare that none of the results described in this manuscript has ever been submitted elsewhere. We hope that the revised manuscript meets all the requirements.

With best wishes,

Qinghua Cui

Department of Biomedical Informatics

School of Basic Medical Sciences

Peking University, Beijing, China

Reviewer #1 (Remarks to the Author):

Comments to the Author

This manuscript presents i-Fold, a modified AlphaFold2-based framework that incorporates residue importance scores (RIS) derived from ESM-2 into the training loss. The idea of guiding structure prediction using residue-level importance is interesting, and the authors provide benchmark and independent dataset evaluations showing improved performance compared with AlphaFold2. Overall, the manuscript is clearly written, and the study is potentially useful to researchers working on protein structure prediction.

I have several suggestions that may help improve clarity, presentation, and the completeness of the evaluation.

1. Fig. 1 would benefit from a clearer illustration of the overall pipeline. It is currently, RIS are generated, where they are introduced into the model, and which components they influence during prediction. A revised schematic that explicitly shows the RIS module and its points of interaction with the AlphaFold2 pipeline would improve readability and make the method easier to follow.

Response:

Thank you for your suggestion. We have redrawn Figure 1, and in the updated figure, we added the calculation process of RIS and its impact on the loss function during training to make it clearer.

2. The manuscript evaluates prediction accuracy mainly using RMSD. While RMSD is informative, it is sensitive to local structural deviations and protein length. Standard structure prediction studies typically report additional metrics such as TM-score or IDDT. Including these metrics would provide a more comprehensive assessment.

Response:

We have supplemented the evaluation of TM-score and IDDT (Fig. 3D, E) and added the specific implementation details in the “Model performance evaluation” section.

3. Currently, the method is compared only with AlphaFold2. If possible, additional comparisons with more recent structure prediction methods, such as AlphaFold3, Chai-1, Boltz-1, etc., would better position the contribution of i-Fold and strengthen the conclusions.

Response:

Since the model weights of AlphaFold3 are not open-source, making deployment relatively difficult, we have supplemented the performance comparison of i-Fold with the open-source methods Chai-1 and Boltz-1 in our paper. The results show that i-Fold achieves performance comparable to both methods. Although limited by current computational resources, we were unable to use optimal hyperparameters or perform more detailed fine-tuning during training, resulting in i-Fold not significantly outperforming the two methods. However, we wish to emphasize that the main contribution of i-Fold is not to refresh absolute accuracy by stacking parameters or using enormous computational power, but rather to propose a "model-free" RIS-weighted training strategy. This strategy, through an improved loss function, guides the model to learn functionally critical regions more accurately. As we mentioned in the discussion section, the RIS weighting strategy is a highly versatile "hot-pluggable" module. Our experiments have demonstrated its significant gains on the AlphaFold2/OpenFold architecture. Theoretically, this strategy can be easily integrated into the loss functions of newer architectures such as Chai-1 or Boltz-1. We believe that with more sufficient computational support, this weighting bias mechanism will further unlock the potential of these large-scale models.

4. The training procedure is described only superficially. Key details such as batch size, number of steps or epochs, recycling settings, and number of trained model parameter sets are not provided. The description of training duration also appears incomplete. Dataset preparation is similarly unclear: although the authors state that 185,158 structures were used for training and 3,599 sequences for testing, it is not specified whether redundancy filtering or homology-based separation was applied. Without this information, potential data leakage cannot be ruled out.

Response:

In our revised manuscript, we have supplemented the description of the training details (see “Model training” section).

Regarding the benchmark test set consisting of 3,599 protein structures, we would like to clarify that these sequences were indeed randomly sampled from the initial dataset, and no strict homology-based deduplication was performed against the remaining 181k training structures. The main reason for constructing this dataset was to establish a tightly controlled baseline comparison—ensuring that both the AlphaFold2 and i-Fold models were trained and evaluated under exactly the same data conditions (thus, we did not evaluate the newly added Chai-1 and Boltz-1 on this dataset). Although we cannot deny that the absolute prediction accuracy might benefit from sequence homology, the relative performance improvement (r.m.s.d. reduced from 1.794 Å to 1.482 Å) provides a robust and fair metric that demonstrates the architectural advantage of the RIS weighting strategy. Subsequently, to rigorously evaluate the model’s generalizability on completely new data, we constructed an independent test set of 167 newly released protein structures, which completely eliminates the risk of data leakage.

5. When comparing iFold and AlphaFold2 modeling performance, please clarify whether the number of generated samples (or recycling/sampling strategies) is controlled to be consistent across methods. If not already stated, the authors should explicitly describe the sampling protocol to ensure the fairness of the comparisons.

Response:

Thank you for your constructive criticism. We explicitly confirm that the recycling /sampling strategy was strictly controlled to be identical between our proposed i-Fold and the baseline AlphaFold2. Specifically, during the prediction phase on all test sets, both models were configured to use 3 recycling iterations, and for each target protein, a single structure sample was generated using a single set of training parameters per method. For neither model did we employ any large-scale sampling or model ensembling strategies (which are known to artificially boost performance metrics). Additionally, we evaluated unrelaxed structures directly to assess the raw neural

network outputs, without the confounding smoothing effects of a force field. Since the sampling space and generation budget were kept completely constant, the improvements observed in i-Fold can be attributed solely to the architectural innovation of our RIS-weighted loss function.

6. Additional ablation experiments would substantially strengthen the methodological claims. In particular:

- (1) A control experiment using randomized RIS weights (if computationally feasible) to quantify the real contribution of RIS.
- (2) A clear comparison of different training strategies, e.g., structure-module-only fine-tuning vs. full-model retraining, if the current model involves full retraining.

Response:

(1) We have added an experiment: training the model with randomly shuffled RIS weights. The results are shown in Figure 2D, where the median r.m.s.d. increased to 1.988 Å, which is even higher than the 1.794 Å achieved by AlphaFold2. This result confirms the effectiveness of the RIS weighting strategy; when the weights are randomly assigned, they become “evolutionary noise” that misdirects the model’s attention, thereby hindering accurate learning of protein folding patterns.

(2) In our current study, we deliberately chose full model training from scratch as our primary approach for two reasons. First, the core innovation of i-Fold is the RIS-weighted FAPE loss. If we froze Evoformer and only fine-tuned the structure module, the evolutionary signals encoded in the RIS weights would only adjust the final coordinate projection. Full model training allows gradients from the FAPE_{RIS} loss to backpropagate through the entire network, ensuring that the complex attention maps within Evoformer learn to intrinsically prioritize functionally critical residues during feature extraction. Second, in practice we found that fine-tuning requires substantially more computational resources than training, mainly because the structural violation loss computed during fine-tuning is quite expensive. Given our current computational budget, it was difficult to perform model fine-tuning. Nevertheless, we consider this suggestion beneficial. In the Discussion section, we note that further applying the RIS

strategy to larger-scale models and more extensive fine-tuning represents a very valuable direction for future research.

7. The explanation for the poor performance on target 7JW6 relies on the concept of "evolutionary noise" generated by ESM. As this is a singular anecdotal example, the manuscript would benefit from a quantitative analysis determining the frequency of such RIS misweighting events. Furthermore, the authors should investigate if there is a systematic correlation between performance degradation and anomalous entropy distributions. Furthermore, the discussion could be strengthened by analyzing the underlying causes of these failures and proposing concrete improvement strategies, possibly accompanied by preliminary validation experiments.

Response:

Thank you for your suggestion. We have added a systematic analysis of evolutionary noise in the Results section. Specifically, we analyzed the protein cases in the independent test set where i-Fold underperformed compared to AlphaFold2, examining the correlation between their r.m.s.d. and evolutionary noise. Here, we defined the Anomalous RIS Ratio (ARR), i.e., the ratio of the average RIS in ordered regions to that in disordered regions. This metric can be used to measure the level of evolutionary noise: under normal circumstances, ARR should be less than 1, and a larger ARR indicates a more anomalous distribution of RIS weights. Finally, we found that the Spearman correlation coefficient between ARR and the r.m.s.d. of i-Fold predictions was 0.43, and the correlation coefficient between ARR and the r.m.s.d. increase of i-Fold relative to AlphaFold2 was 0.56 (Supplementary Fig. S6). This indicates that the prediction performance of i-Fold largely depends on the quality of RIS. We have further discussed future directions for improvement on this issue in the Discussion section.◦

8. There is a conflicting report regarding the benchmark dataset size. The Abstract cites 3,559 structures, whereas both the Results and Methods sections refer to 3,599 sequences. Please ensure these figures are consistent throughout the text.

Response:

Thank you for your correction. We have revised the typo in the abstract; it should be 3599 instead of 3559.

9. In Fig. 2C, the axis labels (“Training epoch” and “Validation Loss”) follow inconsistent capitalization conventions. The authors may consider standardizing figure label formatting and checking other figures for consistency.

Response:

We have corrected the axis labels of Fig. 2C to maintain a consistent uppercase format.

10. Some quantitative axes appear to lack explicit units (e.g., Fig. 3C). For measurements such as structural deviation, indicating the unit would help readers interpret the scale of the results.

Response:

We have revised Figure 3 and added the axis titles and corresponding captions that were missing in the original figure to make it clearer.

11. Several figures are visually coarse and do not clearly convey the methodology. For example, the overview in Fig. 1 does not clearly illustrate the computational pipeline or where RIS weighting is applied. Improving figure clarity would enhance the readability of the manuscript.

Response:

We have redrawn Figure 1 and reviewed all other figures in the text to ensure that the presented results are clear and easy to understand.

12. Several spelling errors throughout the manuscript and supplementary files should be addressed to improve professional clarity. Specifically, please correct "optimazed" to "optimized" and "wighting" to "weighting" in the Methods section (Page 10), and fix "Funtion" to "Function" in the title of Supplementary File S5.

Response:

Thank you for your thorough review. We have corrected the aforementioned spelling

errors.

Reviewer #2 (Remarks to the Author):

The authors present i-Fold, an adaptation of the AlphaFold2 architecture that integrates residue-specific importance scores (RIS) derived from the ESM2 protein language model into the FAPE loss function. The manuscript's primary strength lies in its "simple yet effective" approach—achieving performance gains without adding new parameters or significantly increasing computational overhead. However, several critical methodological ambiguities regarding the training process, data controls, and the nature of the RIS metric must be addressed.

Major Revisions

1. In the section Methods-Model training, the authors state that models were "trained on four NVIDIA A100 GPUs... completing 10 over approximately seven days".

Is that 10 epochs? The author must clarify this sentence.

Original AlphaFold2 training required massive TPU resources over several weeks. Did the authors initialize the model with weights from a pre-trained OpenFold/AlphaFold2 checkpoint? If this is a fine-tuning exercise rather than a "retraining" from scratch as implied, the manuscript must be updated to reflect this, as the 10-epoch gain might represent localized optimization rather than a fundamental architectural shift.

Response:

We sincerely thank the reviewer for pointing out this typographical error and for raising critical questions regarding our training methodology and computational scale.

First, "10" indeed refers to "10 epochs." We have revised the sentence in the manuscript to read: "...completing 10 epochs over approximately seven days."

Regarding the model training question, we would like to clarify that we did not initialize our models with pre-trained weights from OpenFold or AlphaFold2 checkpoints. Both our proposed i-Fold model and our baseline AlphaFold2 model were trained entirely from scratch with random parameter initialization. The original AlphaFold2 training

used substantial TPU resources over several weeks, not only because it utilized 181k PDB structures but also because it involved a large-scale self-distillation dataset and ran for hundreds of thousands of steps. Given our computational limitations, it was infeasible to replicate the exact original AlphaFold2 training; therefore, we adopted a strictly controlled ablation strategy. As described in the added “Baseline methods” section, we retrained a baseline AlphaFold2 model from scratch using the OpenFold framework, under exactly the same hardware, dataset, and training duration as i-Fold. Thus, the performance improvements observed in our study come from comparing two models that started from the same random initialization and received the exact same constrained training regimen. The fact that i-Fold achieved significantly lower prediction error than the retrained AlphaFold2 baseline within this limited 10-epoch window demonstrates that our RIS-weighted loss function provides a fundamental architectural advantage. Applying this strategy directs the model’s attention to functionally critical and evolutionarily constrained regions from the very beginning of the learning process, enabling better representation and markedly faster convergence.

2. The manuscript does not explicitly state whether structural templates were used during the training and prediction phase for i-Fold and the AlphaFold2 baseline. If templates were used, the observed improvements might stem from template-driven modeling rather than the RIS-weighted loss.

Response:

We sincerely thank the reviewer for pointing out this omission. We have added the usage instructions for templates in the Methods section. During training and inference, we followed the standard OpenFold/AlphaFold2 protocol, enabling up to 4 templates per target as input (`max_templates=4`), and used a standard PDB template database with appropriate release date cutoffs to prevent test set leakage. We would like to emphasize that during both training and prediction, our baseline AlphaFold2 and the proposed i-Fold model were provided with exactly the same multiple sequence alignments (MSAs) and exactly the same structural templates. Since the availability of templates remains completely unchanged between the two models, any relative performance

improvement—such as the substantial reduction in median r.m.s.d. observed for i-Fold—can be attributed solely to our novel RIS-weighted loss function, rather than to differences in template availability.

Furthermore, our analysis of low-homology sequences in the paper further confirms that the performance gains of i-Fold are not template-driven (Supplementary Fig. S5). By definition, these “orphan” proteins lack high-quality homologous templates in the PDB. If the success of i-Fold were primarily template-driven, its performance on these targets would collapse to the level of AlphaFold2. However, as our results show, on these specific targets, i-Fold achieved a 1.08 Å reduction in mean r.m.s.d. compared to AlphaFold2. This strongly demonstrates that the RIS weighting strategy, by leveraging the deep representations of the ESM language model, effectively compensates for sparse evolutionary and template signals, thereby yielding genuine architectural improvements independent of template constraints.

Please clarify whether structural templates were utilized in this work. Additionally, which specific version of AlphaFold2 (e.g., v2.0 vs v2.3) was used for comparison, and were the same MSA/template inputs provided to both models to ensure a fair comparison? Furthermore, please specify the number of models generated per target for both i-Fold and AlphaFold2. Was the model selected for evaluation the 'best' among these predictions, and if so, what were the specific criteria (e.g., RMSD) used for this selection?

Response:

Since we needed to modify the loss function of AlphaFold in our experiments, and AlphaFold2 itself does not open-source its training code, we used OpenFold as our implementation framework. OpenFold is a reproduction based on AlphaFold v2.0, so the term AlphaFold2 in this paper refers to version 2.0. In our experiments, both AlphaFold2 and i-Fold were provided with the same number of MSAs and the same number of templates to ensure a fair comparison. Moreover, we allowed the model to generate only one prediction per target in our experiments, thus no result selection was involved; the same applies to the newly added Chai-1 and Boltz-1 methods in this

revision. Incidentally, if one needs to select the “best” among multiple model predictions, the evaluation metric should be pLDDT, because in practical applications the r.m.s.d. cannot be known in advance.

3. In section ‘RIS improves protein structure prediction through multiple mechanisms’, the authors argue that RIS captures information beyond evolutionary conservation, citing the protein HGSNAT (8VKJ) protein as an example where AlphaFold2 failed despite being a "high-homology" sequence. Do the authors use "raw hits" to reflect MSA diversity? Raw hits can be highly redundant and do not accurately reflect the Neff (Number of effective sequences). A target with 10,000 hits might still have a very low Neff, leading to AF2 failure.

Please provide the specific Neff values for the 8VKJ target. Furthermore, provide a formal definition of the "MSA-derived conservation score" used in the correlation analysis in Figure 5B to ensure the 0.45 Spearman's Rho is a meaningful baseline.

Response:

Thank you for your constructive criticism. All mentions of “MSA hits” in our manuscript refer to effective MSA hits, and we have emphasized this point in the revised manuscript. Specifically, we used a custom Python script to read the a3m file generated by multiple sequence alignment, clustered the sequences with a similarity threshold of 0.8, and defined the weight of each sequence as 1 divided by the total number of sequences in its cluster. The effective MSA hits were then defined as the sum of the weights of all sequences. For example, the raw a3m file of protein 8VKJ contained 3,535 raw hits; after calculation, the effective MSA hits were 2,578.13. In addition, we have added the detailed definition and calculation formula of the “MSA-derived conservation score” in the Methods section of the manuscript.

4. AlphaFold2 produces a pLDDT score (Predicted Local Distance Difference Test) as a measure of local confidence. Since RIS is intended to highlight functionally and structurally critical residues, there should be a logical relationship between these metrics.

Please provide a correlation analysis between the ESM-derived RIS and the model's final predicted pLDDT scores. If RIS is merely highlighting regions the model already identifies as high-confidence, its utility is redundant; if they are uncorrelated, the authors should explain why forcing the model to focus on "important" but "low-confidence" regions improves overall global accuracy.

Response:

We have supplemented the correlation analysis between RIS and pLDDT and found that the correlation is relatively low (median Spearman correlation coefficient: 0.29). Consistent with this observation, we separately calculated the r.m.s.d. in regions with and without secondary structure for proteins in the independent test set and found that the reduction in r.m.s.d. is not limited to regions with secondary structure—that is, the RIS strategy improves overall global accuracy. We have added a discussion of this phenomenon in the manuscript. We propose that accurate modeling of secondary structure regions provides tighter boundary constraints and better spatial topological guidance for flexible linker regions. By improving the modeling accuracy of core functional regions, a cascading effect emerges that globally enhances the topological soundness of the overall conformation.

Minor Points

In Figure 3, subfigures A, C, D, and E lack axis titles, which makes it difficult to interpret the results.

Response:

We have updated Figure 3 and added the axis titles and corresponding captions that were missing in the original figure to make it clearer.

April 24th, 2026

Re: Manuscript No. COMMSBIO-26-0247A

Dear Dr. Yang,

Thank you for giving us the opportunity to revise our paper. We would like to express our thanks to the reviewers for the constructive and positive comments. In response to the requirements, we have revised the manuscript and adjusted its format accordingly, and we hope that it now meets the publishing standards. If there are any problems during the editing process, please feel free to contact me at any time.

We declare that none of the results described in this manuscript has ever been submitted elsewhere. We hope that the revised manuscript meets all the requirements.

With best wishes,

Qinghua Cui

Department of Biomedical Informatics

School of Basic Medical Sciences

Peking University, Beijing, China

Comment 1:

1. The author acknowledged that the benchmark test set ($n = 3,599$) was randomly sampled without homology-based deduplication, and I agree that the relative improvement (RMSD $1.794 \rightarrow 1.482$ Å) remains informative as a matched-condition comparison.

However, I respectfully disagree with the claim that the independent test set of 167 newly released structures "completely eliminates the risk of data leakage." Release date alone is not a sufficient criterion: newly deposited PDB structures predominantly belong to previously characterized protein families, and their sequences typically have multiple homologs already present in the AlphaFold2 training data and in BFD/UniRef. The standard for a leakage-controlled test set is sequence identity to training data (commonly $< 30\%$ pairwise identity, or cluster-based splitting), not deposition date.

The authors already have the analytical tools required to address this. In the Methods, they define an "effective MSA hits" metric based on 0.8-similarity clustering of a3m alignments, and they apply this metric to the benchmark test set (Fig. 5, Supplementary Fig. S5). However, no equivalent analysis is reported for the independent test set of 167 structures. Reporting the effective MSA hits for each of the 167 independent test sequences would help clarify this.

Response 1:

We sincerely thank you for this rigorous suggestion. We agree with your rigorous standard regarding data leakage: in structural bioinformatics, sequence homology is the gold standard for separating test and training data, and a time cut-off alone does not exclude the presence of homology. We acknowledge that our wording in the previous response was not precise enough. To address the core scientific question, we followed your recommendation and calculated the effective MSA hits for all 167 targets in the independent test set (Supplementary File S4). We then stratified these sequences into two groups using 1,000 effective MSA hits as a threshold (high homology vs. low homology). Subsequently, we compared the performance of i-Fold and AlphaFold2 within each stratum and observed consistent and significant improvements of i-Fold in both groups. These results clearly confirm that the performance gains of i-Fold are

driven by the RIS weighting mechanism, rather than by “data leakage” or mere memorization of sequence families.

Furthermore, while we acknowledge your point about data homology, our primary motivation for constructing the independent test set of 167 proteins was to simulate a prospective, real-world deployment scenario. When researchers use i-Fold in practice, they are predicting newly discovered or uncharacterized proteins that naturally share homology with the existing protein universe (which is precisely the theoretical basis for AlphaFold2’s use of MSAs as additional input). The fact that i-Fold significantly outperforms the baseline on this completely new, prospective dataset demonstrates its robust practical utility in real-world structural biology workflows.

Comment 2:

2. Assuming the reported test results are obtained on samples held out independently from training, I largely agree with the authors' conclusion that i-Fold's significantly lower prediction error than the retrained AlphaFold2 baseline within the 10-epoch window supports the value of the RIS-weighted loss function.

That said, I would recommend that the authors add an estimate of the best performance i-Fold could be expected to achieve, based on the original OpenFold paper (Ahdriz et al., 2024, Nature Methods 21:1514–1524, doi: 10.1038/s41592-024-02272-z). OpenFold reports that validation IDDT-C α reaches ~ 0.83 within a few thousand training steps at an effective batch size of 132 per step, whereas i-Fold was trained for 12,500 steps (10 epochs $\times$ 1,250 steps, batch size = 8). OpenFold's "few thousand steps" (e.g., 5,000 steps $\times$ 132 = 660,000 samples) corresponds to approximately 40–90 epochs of i-Fold training under the current configuration. Including this reference point would help readers benchmark i-Fold's performance against the expected training trajectory.

Response 2:

We sincerely thank you for this highly constructive and encouraging suggestion. In fact, we mentioned in our previous discussion that, because our training from scratch was limited to a relatively early stage of the learning curve, the model had not yet reached

its full convergence ceiling. You understood our intention clearly and suggested that we provide a specific explanation. Following this excellent suggestion, we have added an estimate of the performance ceiling of i-Fold in the Discussion section.

We now cite the OpenFold paper, noting that the current i-Fold—both in terms of training data scale and training time—is still in the ‘early training stage’ as defined by OpenFold. Given the report that OpenFold achieves 90% of its final model performance early in training, we reasonably estimate that under a training budget comparable to that of OpenFold, the absolute performance of i-Fold would improve by at least 10%, which would be sufficient to make i-Fold a state-of-the-art model.

Comment 3:

3. I appreciate the authors' transparency regarding the relatively low correlation between RIS and pLDDT (median Spearman correlation coefficient: 0.29). However, I believe the response would be substantially strengthened by addressing two specific points that more directly engage with my original question.

First, the reported correlation ($\rho = 0.29$) is computed across all residues, but it would be informative to know whether this correlation differs between the two structural categories the authors themselves analyze. Specifically, what is the RIS–pLDDT correlation within secondary-structure regions versus within non-secondary-structure (flexible linker) regions? A meaningful difference between these two strata would clarify whether RIS provides distinct information primarily in one region type versus the other, and would lend further mechanistic support to the proposed "cascading effect."

Second, the core question — why forcing the model to focus on important but low-confidence regions improves overall global accuracy — is not directly addressed by the per-region RMSD comparison ($2.67 \rightarrow 2.13 \text{ \AA}$ for secondary-structure regions, $2.80 \rightarrow 2.48 \text{ \AA}$ for non-secondary-structure regions). To directly support the authors' claim that RIS improves global accuracy beyond regions of secondary structure, I recommend adding a concrete example: identify one or more residues (or contiguous regions) in the independent test set that have high RIS but low AlphaFold2 pLDDT, and demonstrate

that i-Fold improves the local prediction accuracy at these specific positions. Such a case study — ideally with structural visualization — would provide the mechanistic intuition that a single correlation coefficient cannot convey, and would directly answer the reviewer's question of why redirecting attention to "important but uncertain" residues yields global rather than merely local improvements.

Response 3:

We are deeply grateful to you for these suggestions.

In response to the first question, we followed ‘your suggestion and performed a comprehensive stratified analysis of pLDDT and RIS. Specifically, we calculated the correlation between RIS and pLDDT from both AlphaFold2 and i-Fold. The results showed that RIS exhibits low correlation with pLDDT regardless of the source and region (median correlation coefficient: AlphaFold2: total 0.27, structured 0.23, non-structured 0.31; i-Fold: total 0.29, structured 0.29, non-structured 0.31). We believe this finding confirms that RIS is semantically orthogonal to model “confidence”, but it does not clearly explain what specific role the information provided by RIS at different regions plays. In fact, we observed that in most proteins, only the first one or two residues at the N-terminus and the last one or two residues at the C-terminus exhibited notably lower pLDDT, while the pLDDT at the remaining positions was generally high. We consider this to be a reason why the correlation analysis results are not very definitive. Therefore, we further supplemented the analysis of the median pLDDT differences between regions with and without secondary structure. The results were similar to those for r.m.s.d.: for AlphaFold2, the median pLDDT values in structured and unstructured regions were 87.28 and 83.61, respectively, while i-Fold increased these to 93.01 and 88.99 (Supplementary File S6). We believe that the universal improvement in pLDDT across both regions provides additional evidence for the “cascading effect”.

Regarding the second question, we believe that protein 8ida serves as a good example (Supplementary Fig. S7). In this case, AlphaFold2 exhibited a high prediction r.m.s.d. of 12.86 Å, primarily because it mispredicted the N-terminal alpha helix of the protein. This error triggered a cascade effect, leading to incorrect predictions for the flexible

linker region after the alpha helix. Accordingly, AlphaFold2 gave low pLDDT scores in this region. In contrast, RIS significantly emphasized the alpha helix in this region, resulting in a largely correct protein conformation. Thus, accurate modeling of specific secondary structure regions also helps improve the modeling accuracy of their surrounding flexible linker regions.

June 23rd, 2026

Re: Manuscript No. COMMSBIO-26-0247B

Dear Dr. Yang,

Thank you for giving us the opportunity to revise our paper. We would like to express our thanks to the reviewers for the constructive and positive comments. In response to the requirements, we have revised the manuscript and adjusted its format accordingly, and we hope that it now meets the publishing standards. If there are any problems during the editing process, please feel free to contact me at any time.

We declare that none of the results described in this manuscript has ever been submitted elsewhere. We hope that the revised manuscript meets all the requirements.

With best wishes,

Qinghua Cui

Department of Biomedical Informatics

School of Basic Medical Sciences

Peking University, Beijing, China

Comment 1:

The authors have addressed most of my comments. I only have one minor suggestion regarding terminology.

I suggest avoiding the term "low-homology group" for sequences with effective MSA hits < 1000 , as this threshold does not necessarily indicate low homology in the conventional sense. A more precise description would be a medium-homology group, which would more accurately reflect what was actually measured.

Response 1:

We sincerely thank the reviewer for pointing out this terminology issue. We agree that "medium-homology" is a more precise and accurate description for sequences with effective MSA hits < 1000 . We have revised the manuscript and replaced all instances of "low-homology" with "medium-homology" when referring to this specific group.