Reciprocal Best Matching: a new pipeline for scoring models with unknown stoichiometry in CASP experiments

Rongqing Yuan
UT Southwestern Medical Center
5323 Harry Hines Blvd. Dallas, TX, 75390
rongqing.yuan@utsouthwestern.edu

Jing Zhang
UT Southwestern Medical Center
5323 Harry Hines Blvd. Dallas, TX, 75390
jing.zhang@utsouthwestern.edu

Qian Cong
UT Southwestern Medical Center
5323 Harry Hines Blvd. Dallas, TX, 75390
qian.cong@utsouthwestern.edu

Abstract

Accurate prediction of protein complex structures remains a significant challenge, particularly when stoichiometry information is unavailable. In the recent Critical Assessment of Structure Prediction Round XVI (CASP16), the “Phase 0” challenge was introduced to stimulate progress in this area. However, existing evaluation tools, such as OpenStructure, introduce systematic biases when evaluating models with stoichiometries different from the target, sometimes favoring those with excess subunits and inflating scores for incorrect stoichiometries. To address this issue, we developed the Reciprocal Best Matching (RBM) pipeline. RBM compares predicted and target structures by bidirectionally matching interfaces and assigning penalizations to unmatched interfaces. This approach penalizes incorrect stoichiometries in a consistent and unbiased manner while preserving strong correlation with established CASP metrics. Application of RBM in CASP16 assessments revealed improved discrimination between correctly and incorrectly stoichiometric models. We also provide a standalone software for our RBM pipeline, which is useful for protein complex structure prediction evaluation and future CASP experiments.

1 Introduction

In contrast to the continual success of monomeric protein structure prediction in the recent Critical Assessment of Structure Prediction Round XVI (CASP16) experiment, the problem of protein complex structure prediction remains a largely unsolved challenge [17, 18]. Traditional CASP experiments ask participants to predict protein complex structures given the knowledge of the stoichiometry of the complex. Even in this simplified scenario, the success rate in this category is about 50%. Furthermore, for predictions to be truly powerful in guiding and potentially replacing experimental characterizations, it is essential to predict the 3D structure of protein complexes without knowing stoichiometry in advance.

To stimulate further progress in this field, CASP16 introduced a new challenge, referred to as the “Phase 0” challenge [7, 9], whereas the traditional CASP complex prediction challenge is referred to as “Phase 1”. The targets in Phase 0 comprised a subset of Phase 1 targets, and the predictors were required to predict the protein complex structures without stoichiometry information. Additionally, in

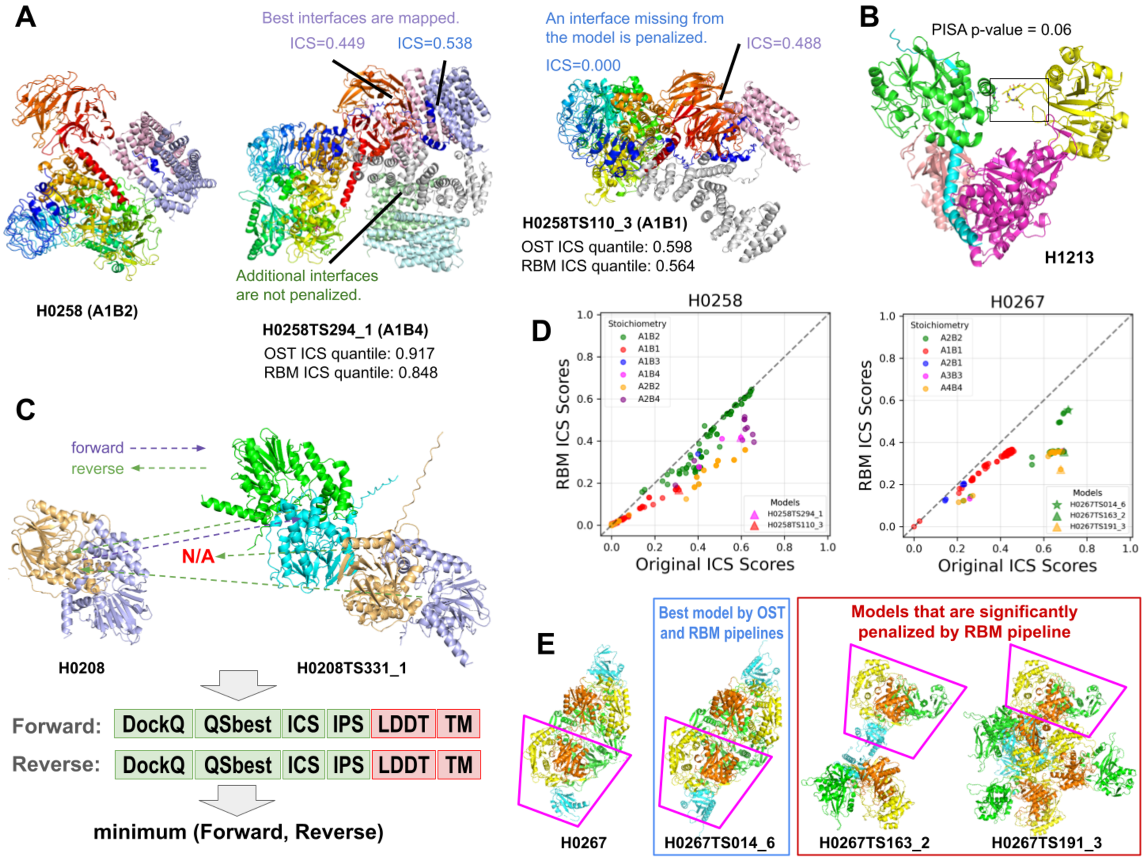

Figure 1: (A) A systematic bias exists in the original CASP16 assessment method: predicting fewer subunits is penalized more than predicting more subunits. (B) Small interfaces can be biologically meaningful. (C) Graphic illustration of RBM approach. (D) Example of ICS scores computed using OST software and RBM pipeline. Models with the correct stoichiometry are colored in green. RBM pipeline is more in favor of the targets that have the correct stoichiometry. (E) Target H0267 and several models: blue box, best model (labeled as a star in D); red box, models showing the largest difference in OST and RBM scores (labeled as triangles in D). The structures are colored by domains instead of by chains, the building block (in magenta trapezoid) is correctly predicted in all these models, but they are assembled incorrectly in the two models (in red box) strongly penalized by our RBM pipeline.

Phase 1, stoichiometry information was unavailable for certain filamentous targets, and the predictors should predict the repetitive structural unit of the filaments. By incorporating this additional step of stoichiometry prediction, the challenge encouraged the development of more comprehensive modeling pipelines. Such efforts will benefit the structural biology community, because obtaining stoichiometry information for protein complexes is not trivial [15, 20, 3]. Furthermore, predicted stoichiometry and protein complexes can assist the interpretation and validation of experimental structures. For example, it remains a challenging problem to identify biological assemblies from crystal contacts [5, 4], and resolved crystal structures often do not contain the assembly under physiological conditions [8].

In CASP experiments, most complex assessment scores are calculated using a software package called OpenStructure (OST) [2]. The scores commonly used by CASP assessors can be broadly divided into two categories: those that focus on the protein-protein interfaces and those that evaluate the quality of the entire structures. The former, such as ICS, IPS, QSbest [10], and DockQ [1], evaluate the quality of interfaces between interacting chains. The latter, such as lDDT [12] and TM-score [19], measure the overall structural similarity between the predicted structure and the reference target. These scores have been widely used in previous CASP assessments [10, 7, 13, 14].

OST attempts to find a one-to-one match between chains in the target and chains in a model. This strategy is effective for previous CASP experiments and the Phase 1 challenge in CASP16, where the models are expected to contain the same set of chains as the target. However, we found that this established strategy could unfairly bias towards models with more subunits in Phase 0: OST selects the best subset of chains in a model to maximize its similarity to the target, meaning that predictions with more subunits have higher chances of containing the correct interface (Fig. 1A middle) than those with fewer subunits (Fig. 1A right). In extreme cases, predictors could achieve high interface scores by enumerating multiple possible interfaces between a pair of chains in one model. Although upon our inspection during our assessment, no predictors deliberately exploited this caveat to achieve artificially higher scores, this feature of OST nonetheless introduces a systematic bias.

In addition, we and previous CASP assessors [14] have noticed that the overall interface scores, such as ICS, IPS, and QS best, computed by OST are dominated by large interfaces. This occurs because all interface residues are pooled together to calculate various accuracy metrics, such as precision and recall. However, some small interfaces are biologically meaningful (Fig. 1B) and can be critical for the structure and function of the complex. Furthermore, compared to those large and stable interfaces that frequently correlate with stronger interactions, small interfaces are usually harder to predict due to weaker coevolutionary and physical signals. Evaluating the prediction quality of these important yet difficult interfaces is essential, and increasing their contribution to evaluation scores is therefore desirable.

2 RBM: a new scoring routine for predictions with uncertain stoichiometry

Motivated by the above observations, we developed a Reciprocal Best Matching (RBM) pipeline during our evaluation of CASP16 protein complexes. In this pipeline, all interfaces present in a target are matched to their best corresponding interfaces in a model, and conversely, all interfaces in this model are matched to those in the target. Any interface that appears exclusively in either the model or the target is penalized by assigning a score of 0 (Fig. 1C).

RBM pipeline evaluates one pair of interacting chains at a time. Scores for all chain pairs in the target are subsequently weight-averaged to yield a “forward” score. Meanwhile, the scores for all interacting chain pairs in the model are weight-averaged to obtain a “reverse” score. The minimum between the “forward” and “reverse” scores is taken as the final RBM score. The weight for each chain pair can be adjusted to emphasize important interfaces. In our CASP assessment, we used $\log_{10}(N_{ave})$ as the weight, where N_{ave} is the average number of interface residues between two interacting chains. Compared with the default weighting strategy that pools all interface residues, our strategy upweights small interfaces to reward success in correctly predicting the more challenging parts.

The RBM pipeline takes as input the PDB structures of both the model and the target, along with the JSON output files generated by OST. The OST output files provide the chemical mapping (i.e., the chains with identical sequences) between the target and the model, which is used to match interfaces between the target and the model. RBM does not introduce any additional scores; rather, it is a routine that can be applied with any existing quality scores that can be used to evaluate the quality of a binary protein complex, such as ICS, IPS, DockQ, QSbest, lDDT and TM-score, as used in CASP16[18]. This design allows it to be seamlessly integrated into any existing CASP assessment pipeline. Implemented in Python, RBM is available as a command-line tool for evaluating protein complex structures in cases where stoichiometry information is unavailable. The package allows users to select which scores to apply and to choose among several weighting strategies.

3 Analysis of scores produced by RBM pipeline

In Fig. 1D, we provide an analysis of the scores computed by the RBM pipeline, using ICS as an example. ICS, defined as the F1-score derived from the precision and recall of inter-chain contacts, has been one of the most widely used scores in CASP complex assessments. As shown in Fig. 1D, ICS scores generated by RBM remain highly correlated with those computed by OST, demonstrating that RBM faithfully reflects model quality. At the same time, models with incorrect stoichiometry tend to receive lower scores under RBM pipeline. Notably, RBM and OST scores differ the most for models that predict more subunits than the correct number of subunits. This is related to the fact

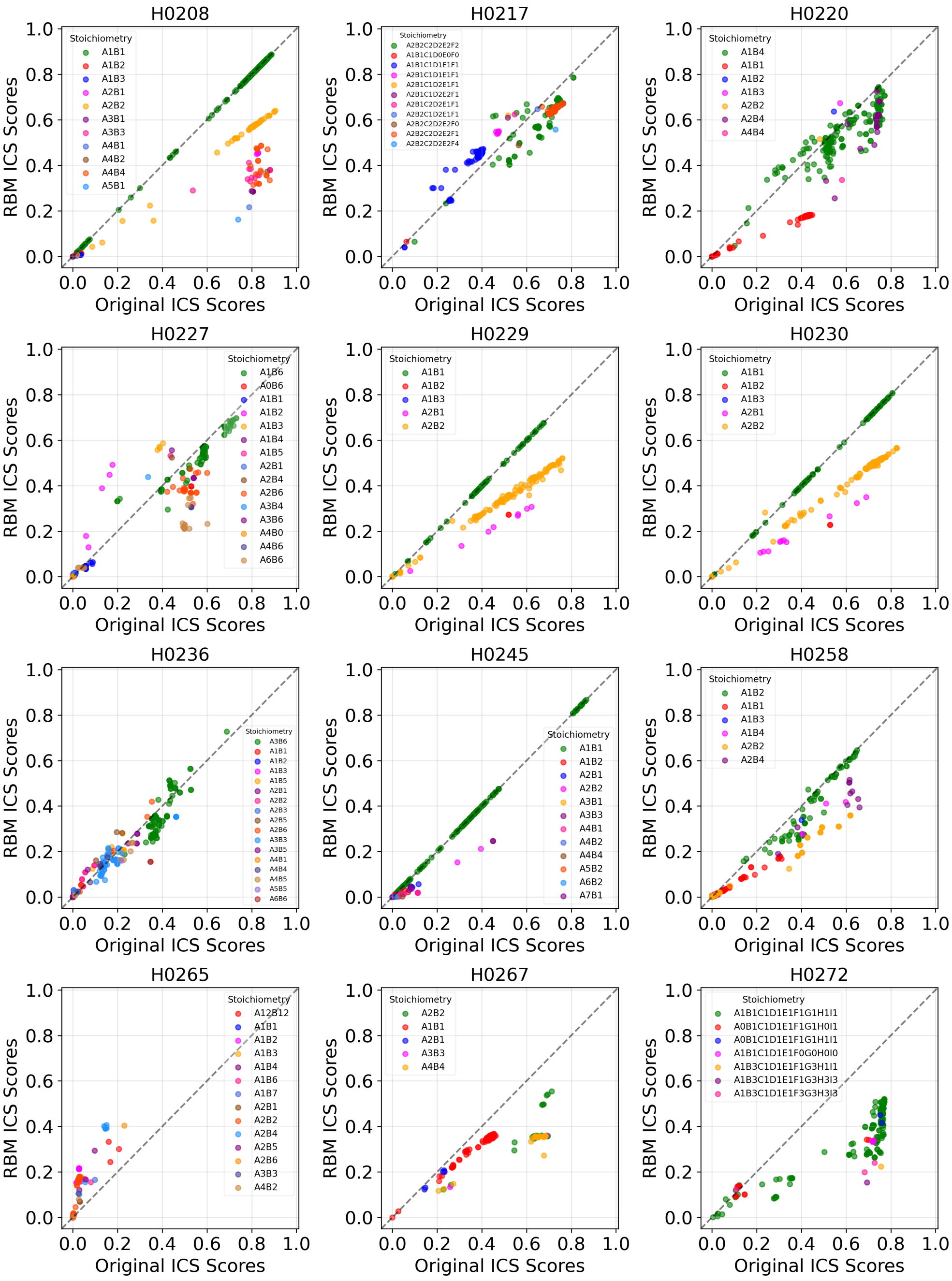

Figure 2: OST score vs. RBM score for Phase 0 hetero-oligomer targets. The minimum of the forward and reverse scores was used as the final model scores. This is the version we decided to use in our complex assessment paper.

that OST might be biased towards such predictions (overprediction of subunits) by focusing on the correctly predicted interfaces, whereas RBM will penalize any wrongly predicted interfaces. In our assessment of CASP16, we initially used a version of RBM pipeline where the forward and reverse scores were averaged (instead of taking the minimum) to obtain the final scores. However, closer inspection of results from this alternative strategy suggests that it might favor models having only the interfaces that are easy to predict. This undesirable bias is clearly revealed for target H0267. H0267 had an A2B2 stoichiometry (Fig. 1E). Under our initial evaluation strategy, we discovered that many A1B1 models received relatively high scores (Fig. 1E and Fig. 3). This occurred because the large A-B interface in H0267 was relatively easy to predict, whereas most predictors that submitted A2B2 models failed to capture the smaller A-A and B-B interfaces. Consequently, A1B1 models were penalized for missing the A-A and B-B interfaces only in the forward direction, whereas A2B2 models with incorrect A-A and B-B interfaces were penalized in both directions (Fig. 1E).

To address the biases revealed by this target, we further adopted a strategy that takes the minimum of the forward score and the reverse score, which penalizes missing interfaces resulting from either incorrect stoichiometry or incorrect prediction. This method yields a distribution of scores shown in Fig. 2, in which models with incorrect stoichiometries — whether over-predicting or under-predicting the number of subunits — are both penalized, which is better aligned with our motivation of developing the RBM score. A third possible strategy is to take the weighted average of all the per-interface scores, regardless of whether the interface is in the target or in the model. While we did not observe a problem with this strategy, a potential concern is that predictors could obtain high scores by repeating a confidently predicted interface many times in the model.

In Fig. 2, Fig. 3, and Fig. 4, we provide the distribution of ICS score under 3 different strategies for all Phase 0 hetero-oligomer targets, excluding antibody-antigen targets[18]. Because no predictors in CASP16 attempt to utilize the potential biases with OST or the original version of RBM to obtain high scores, different strategies do not affect the ranking between CASP16 groups. Nevertheless, we believe our favored strategy, i.e., taking the minimum of the forward and reverse scores, is the most robust. The results for this strategy for ICS are shown in Fig. 2. We considered this strategy the default for the RBM pipeline and used it in this work.

To study the difference between our RBM score and the OST ICS score, we further grouped the models into three categories – correct stoichiometry, more subunits, less subunits – and examined the distributions of each score, as well as the differences between RBM and OST scores (Fig. 5). Overall, RBM applies a stronger penalty to models that contain more subunits than the target. This again indicates that the current OST scores are inflated for such models, and suggests that the RBM pipeline effectively corrects this systematic bias.

Beyond its ability to penalize different models more equitably, the RBM score also remains broadly consistent with the original OST score. For $k = 1, 2, 5, 10, 20$, we count the overlap between OST's top-k models and RBM's top-k models (Table 1). For targets with lower top-k overlap, the discrepancy usually arises because models with incorrect stoichiometry received inflated OST scores but were appropriately penalized by RBM. Such examples include H0229 and H0267 (Fig. 2).

Target	Top-1	Top-2	Top-5	Top-10	Top-20
H0208	0	0	2	3	8
H0217	1	2	2	4	14
H0220	0	0	0	0	6
H0227	1	2	4	9	19
H0229	0	0	0	0	2
H0230	0	0	0	3	9
H0236	1	1	3	3	15
H0245	1	2	5	10	20
H0258	0	0	3	6	9
H0265	0	1	2	8	13
H0267	1	2	2	3	6
H0272	0	1	3	7	12

Table 1: Number of top-k overlapping models between OST ICS and RBM ICS.

We also applied RBM scores to other widely used CASP scores, such as DockQ (Fig. 6) and lDDT (Fig. 7), and compared the values with default OST scores. In the current implementation, we apply a consistent strategy used for interface-based scores: for each interacting pair, we compute the pairwise score and then take a weighted average across all pairs. For DockQ, the RBM-adjusted scores remain highly correlated with the OST scores, likely because DockQ still considers the interface quality. This strategy showed a weaker correlation between the RBM and OST scores for lDDT, suggesting that this pairwise-averaging approach may not be optimal for these alignment-based scores. An alternative solution may be to perform a direct "bidirectional alignment" between the entire model and target structures, rather than averaging over pairwise components.

4 Conclusion

In this work, we introduced RBM, a pipeline designed to provide an unbiased assessment of predicted complex structures with unknown stoichiometry. This pipeline was used to evaluate the Phase 0 challenge in CASP16, which was introduced to stimulate progress in stoichiometry prediction. Given the active community engagement of Phase 0 challenge [6, 11, 16] and the practical need for protein complex structure prediction without prior experimental knowledge, we anticipate a growing need for complex structure evaluation methods that address uncertain stoichiometry. Therefore, we believe the RBM, along with the concept underlying its design, can serve as a good reference for future CASP assessors and method developers in the field. The pipeline is straightforward to use, and the source code with usage instructions is available at https://github.com/RongqingYuan/RBM.

Acknowledgements

This work is supported by the Endowed Scholars Program in UTSW, I-2095-20220331 from Welch Foundation, 1R35GM160468-01 from NIGMS, and NIAID K99AI180984-01A1.The authors acknowledge computing resource support from NSF ACCESS award MED240004, TACC award MCB23014 and NIAIR award NAIRR240466. We thank all CASP organizers, in particular, Andriy Kryshtafovych and John Moult, for their guidance. We also thank Jimin Pei, R. Dustin Schaeffer, Rachael C. Kretsch, Rhiju Das, Jian Zhou, Gabriel Studer, Mark Lensink, and Nick V. Grishin for discussions and help.

References

[1] Sankar Basu and Björn Wallner. DockQ: A Quality Measure for Protein-Protein Docking Models. *PLOS ONE*, 11(8):e0161879, August 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.0161879. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0161879. Publisher: Public Library of Science.

[2] M. Biasini, T. Schmidt, S. Bienert, V. Mariani, G. Studer, J. Haas, N. Johner, A. D. Schenk, A. Philippsen, and T. Schwede. OpenStructure: an integrated software framework for computational structural biology. *Acta Crystallographica Section D: Biological Crystallography*, 69(5):701–709, May 2013. ISSN 0907-4449. doi: 10.1107/S0907444913007051. URL //journals.iucr.org/paper?ic5090. Publisher: International Union of Crystallography.

[3] Gianluca Degliesposti. Probing Protein Complexes Composition, Stoichiometry, and Interactions by Peptide-Based Mass Spectrometry. In M. Cristina Vega and Francisco J. Fernández, editors, *Advanced Technologies for Protein Complex Production and Characterization: Volume II*, pages 41–57. Springer Nature Switzerland, Cham, 2024. ISBN 978-3-031-52193-5. doi: 10.1007/978-3-031-52193-5_4. URL https://doi.org/10.1007/978-3-031-52193-5_4.

[4] Sucharita Dey and Emmanuel D. Levy. PDB-wide identification of physiological hetero-oligomeric assemblies based on conserved quaternary structure geometry. *Structure*, 29(11):1303–1311.e3, November 2021. ISSN 0969-2126. doi: 10.1016/j.str.2021.07.012. URL https://www.sciencedirect.com/science/article/pii/S0969212621002628.

[5] Sucharita Dey, David W. Ritchie, and Emmanuel D. Levy. PDB-wide identification of biological assemblies from conserved quaternary structure geometry. *Nature Methods*, 15(1):67–72, January 2018. ISSN 1548-7105. doi: 10.1038/nmeth.4510. URL https://www.nature.com/articles/nmeth.4510. Publisher: Nature Publishing Group.

[6] Arne Elofsson. AlphaFold3 at CASP16. *Proteins: Structure, Function, and Bioinformatics*, n/a(n/a). ISSN 1097-0134. doi: 10.1002/prot.70044. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.70044. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.70044.

[7] Dmytro Guzenko, Aleix Lafita, Bohdan Monastyrskyy, Andriy Kryshtafovych, and Jose M. Duarte. Assessment of protein assembly prediction in CASP13. *Proteins: Structure, Function, and Bioinformatics*, 87(12):1190–1199, 2019. ISSN 1097-0134. doi: 10.1002/prot.25795. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.25795. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.25795.

[8] Evgeny Krissinel and Kim Henrick. Inference of Macromolecular Assemblies from Crystalline State. *Journal of Molecular Biology*, 372(3):774–797, September 2007. ISSN 0022-2836. doi: 10.1016/j.jmb.2007.05.022. URL https://www.sciencedirect.com/science/article/pii/S0022283607006420.

[9] Andriy Kryshtafovych, Maciej Milostan, Marc F. Lensink, Sameer Velankar, Alexandre M. J. J. Bonvin, John Moult, and Krzysztof Fidelis. Updates to the CASP Infrastructure in 2024. *Proteins: Structure, Function, and Bioinformatics*, n/a(n/a). ISSN 1097-0134. doi: 10.1002/prot.70042. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.70042. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.70042.

[10] Aleix Lafita, Spencer Bliven, Andriy Kryshtafovych, Martino Bertoni, Bohdan Monastyrskyy, Jose M. Duarte, Torsten Schwede, and Guido Capitani. Assessment of protein assembly prediction in CASP12. *Proteins: Structure, Function, and Bioinformatics*, 86(S1):247–256, 2018. ISSN 1097-0134. doi: 10.1002/prot.25408. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.25408. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.25408.

[11] Jian Liu, Pawan Neupane, and Jianlin Cheng. Accurate Stoichiometry Prediction of Protein Complexes by Integrating AlphaFold3 and Template Information, January 2025. URL https://www.biorxiv.org/content/10.1101/2025.01.12.632663v1. Pages: 2025.01.12.632663 Section: New Results.

[12] Valerio Mariani, Marco Biasini, Alessandro Barbato, and Torsten Schwede. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728, November 2013. ISSN 1367-4803. doi: 10.1093/bioinformatics/btt473. URL https://doi.org/10.1093/bioinformatics/btt473.

[13] Burcu Ozden, Andriy Kryshtafovych, and Ezgi Karaca. Assessment of the CASP14 assembly predictions. *Proteins: Structure, Function, and Bioinformatics*, 89(12):1787–1799, 2021. ISSN 1097-0134. doi: 10.1002/prot.26199. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.26199. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.26199.

[14] Burcu Ozden, Andriy Kryshtafovych, and Ezgi Karaca. The impact of AI-based modeling on the accuracy of protein assembly prediction: Insights from CASP15. *Proteins: Structure, Function, and Bioinformatics*, 91(12):1636–1657, 2023. ISSN 1097-0134. doi: 10.1002/prot.26598. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.26598. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.26598.

[15] Carla Schmidt, Christof Lenz, Michael Grote, Reinhard Lührmann, and Henning Urlaub. Determination of Protein Stoichiometry within Protein Complexes Using Absolute Quantification and Multiple Reaction Monitoring. *Analytical Chemistry*, 82(7):2784–2796, April 2010. ISSN 0003-2700. doi: 10.1021/ac902710k. URL https://doi.org/10.1021/ac902710k. Publisher: American Chemical Society.

[16] Wenkai Wang, Yuxian Luo, Zhenling Peng, and Jianyi Yang. Accurate Biomolecular Structure Prediction in CASP16 With Optimized Inputs to State-Of-The-Art Predictors. *Proteins: Structure, Function, and Bioinformatics*, n/a(n/a). ISSN 1097-0134. doi: 10.1002/prot.70030. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.70030. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.70030.

[17] Rongqing Yuan, Jing Zhang, Andriy Kryshtafovych, R. Dustin Schaeffer, Jian Zhou, Qian Cong, and Nick V. Grishin. CASP16 Protein Monomer Structure Prediction Assessment. *Proteins: Structure, Function, and Bioinformatics*, n/a(n/a). ISSN 1097-0134. doi: 10.1002/prot.70031. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.70031. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.70031.

[18] Jing Zhang, Rongqing Yuan, Andriy Kryshtafovych, Rachael C. Kretsch, R. Dustin Schaeffer, Jian Zhou, Rhiju Das, Nick V. Grishin, and Qian Cong. Assessment of Protein Complex Predictions in CASP16: Are we making progress?, May 2025. URL https://www.biorxiv.org/content/10.1101/2025.05.29.656875v1. Pages: 2025.05.29.656875 Section: New Results.

[19] Yang Zhang and Jeffrey Skolnick. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33(7):2302–2309, April 2005. ISSN 0305-1048. doi: 10.1093/nar/gki524. URL https://doi.org/10.1093/nar/gki524.

[20] Susann Zilkenat, Iwan Grin, and Samuel Wagner. Stoichiometry determination of macromolecular membrane protein complexes. *Biological Chemistry*, 398(2):155–164, February 2017. ISSN 1437-4315. doi: 10.1515/hsz-2016-0251. URL https://www.degruyterbrill.com/document/doi/10.1515/hsz-2016-0251/html. Publisher: De Gruyter.

A Appendix / supplemental material

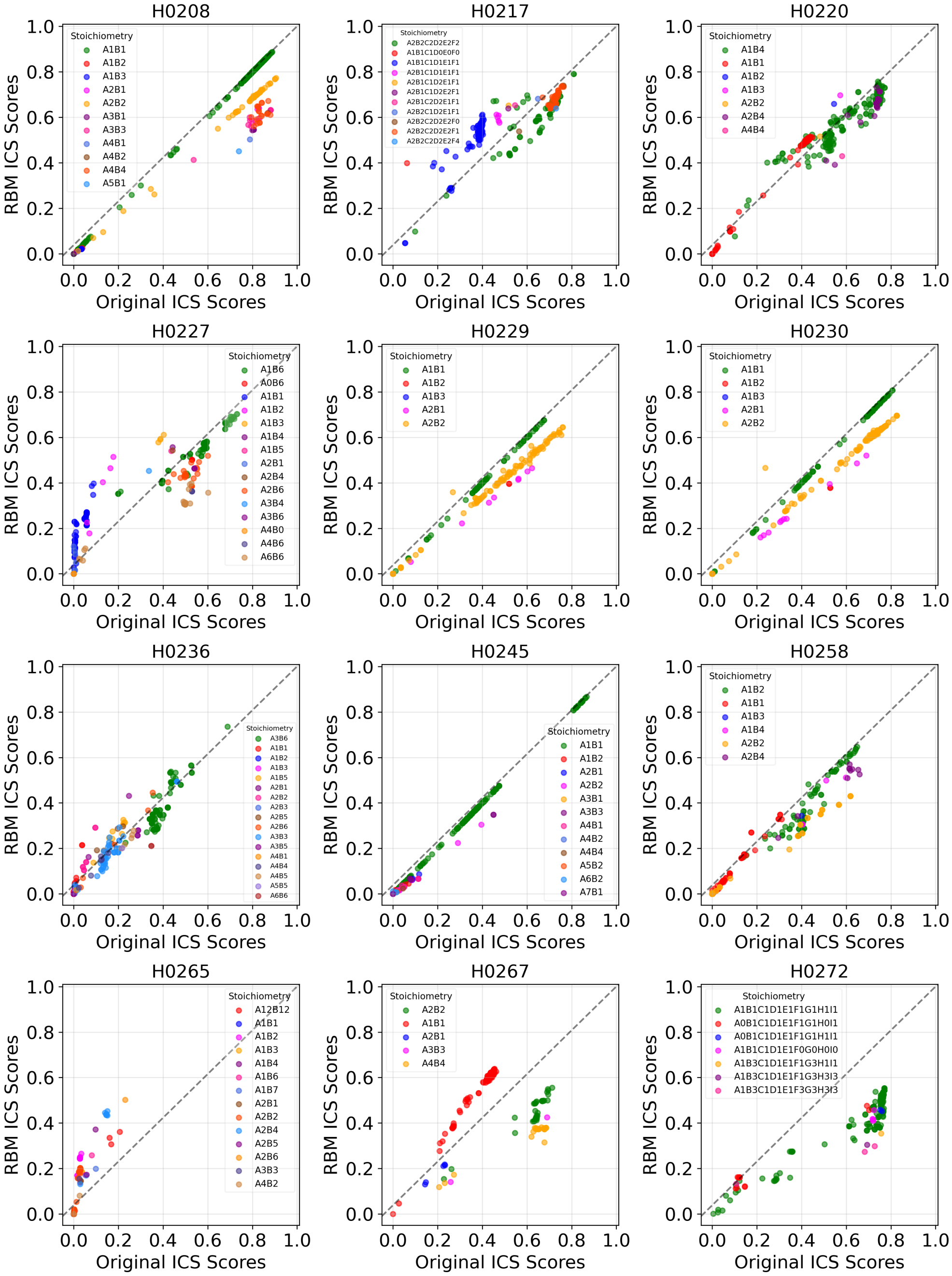

Figure 3: OST score vs. RBM score for Phase 0 hetero-oligomer targets. The forward and reverse scores were averaged to obtain the final model scores. This is the version we initially used in CASP assessment.

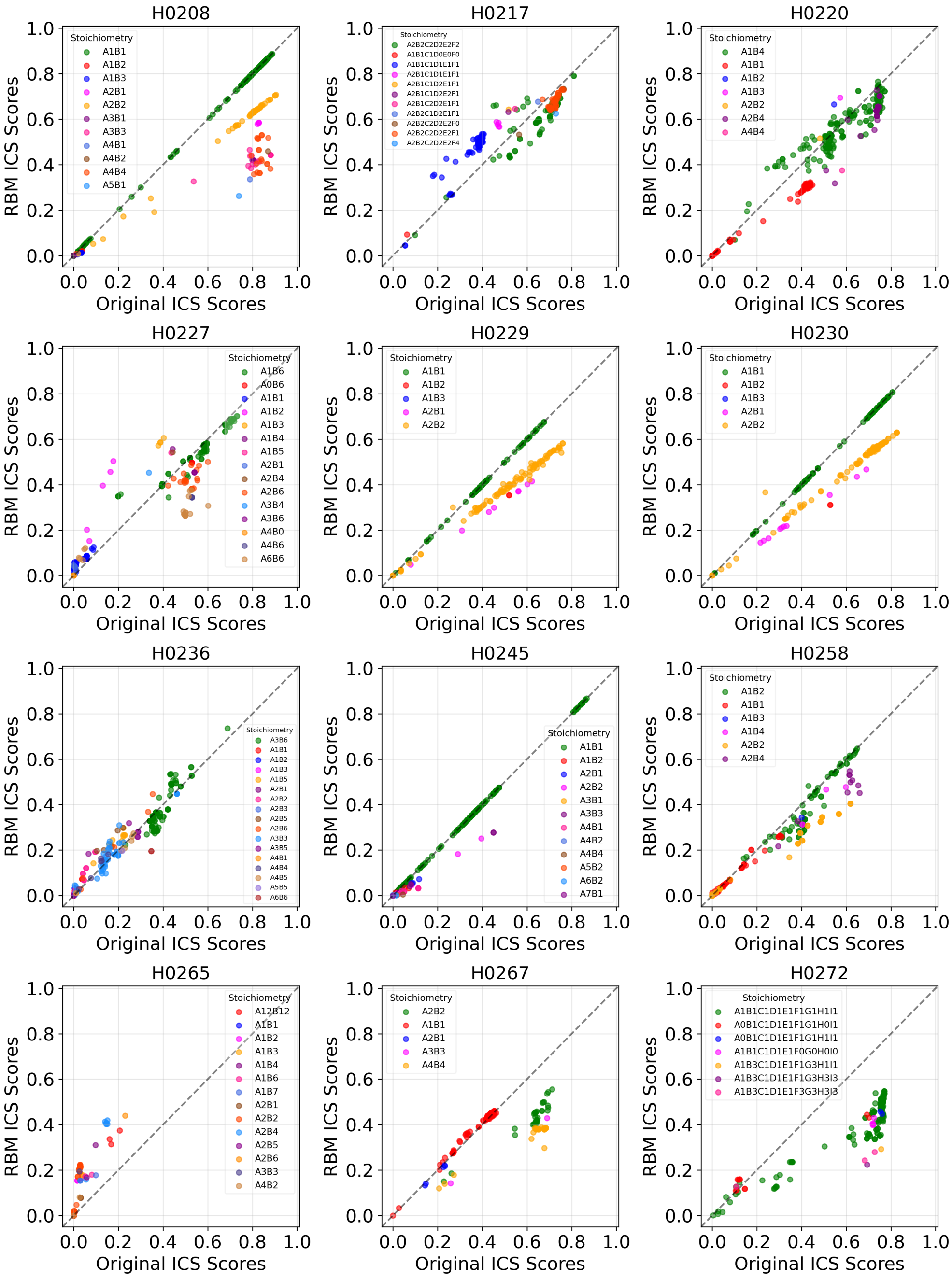

Figure 4: OST score vs. RBM score for Phase 0 hetero-oligomer targets. The final model score is the weighted average of all the per-interface scores, regardless of whether the interface is in the target or in the model.

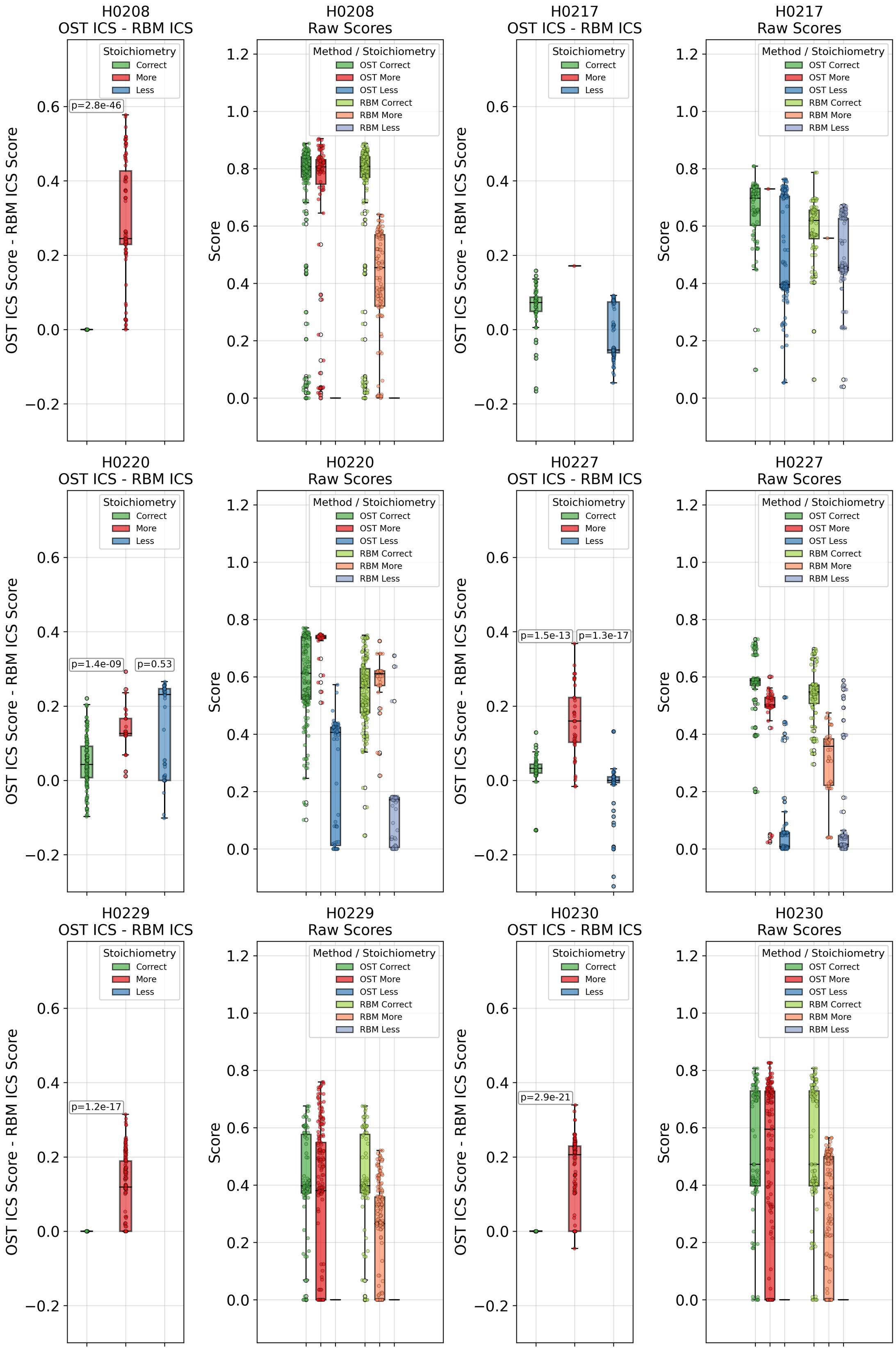

Figure 5: OST score vs. RBM score for Phase 0 hetero-oligomer targets.

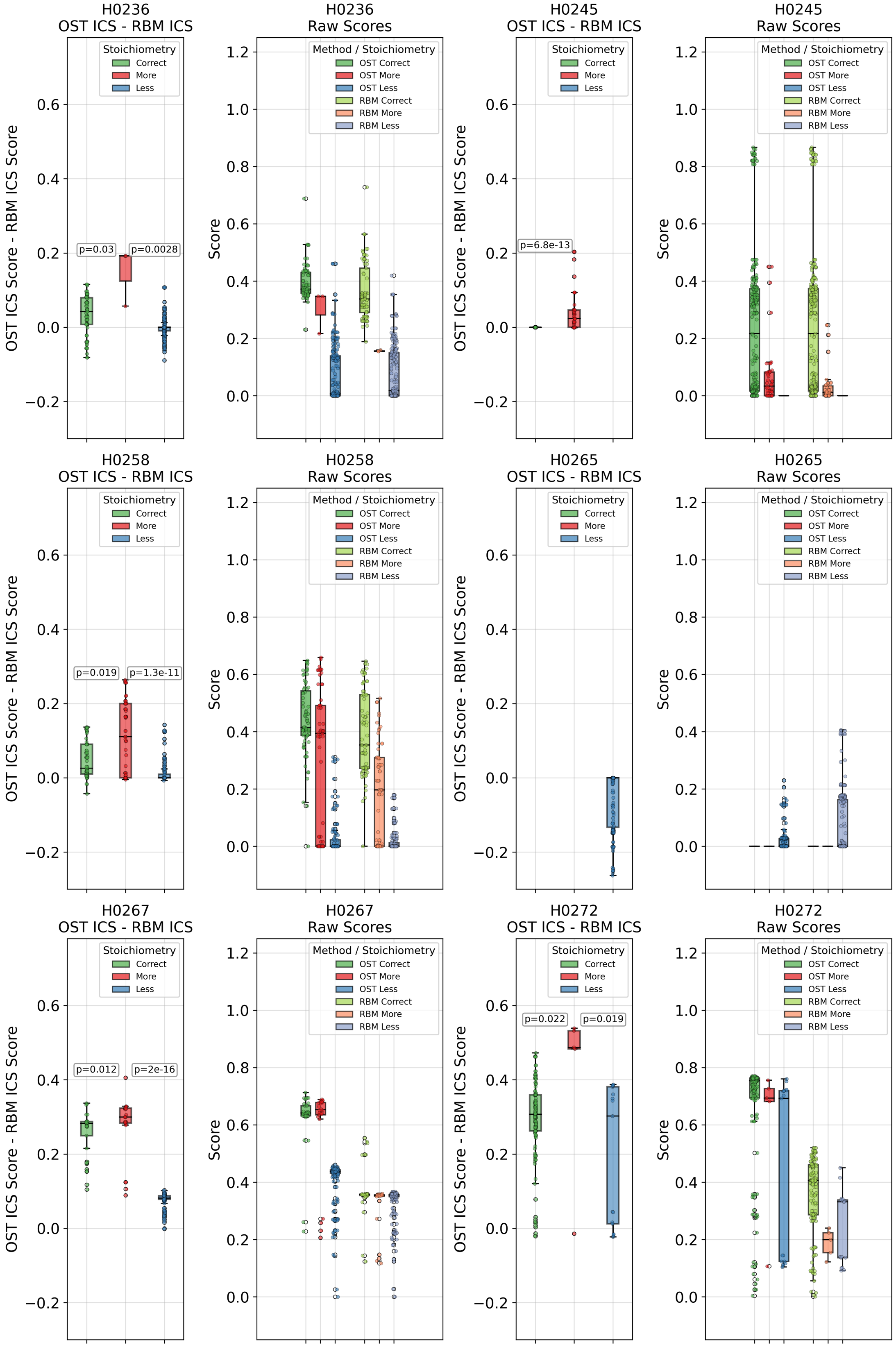

Figure 5: (continued)

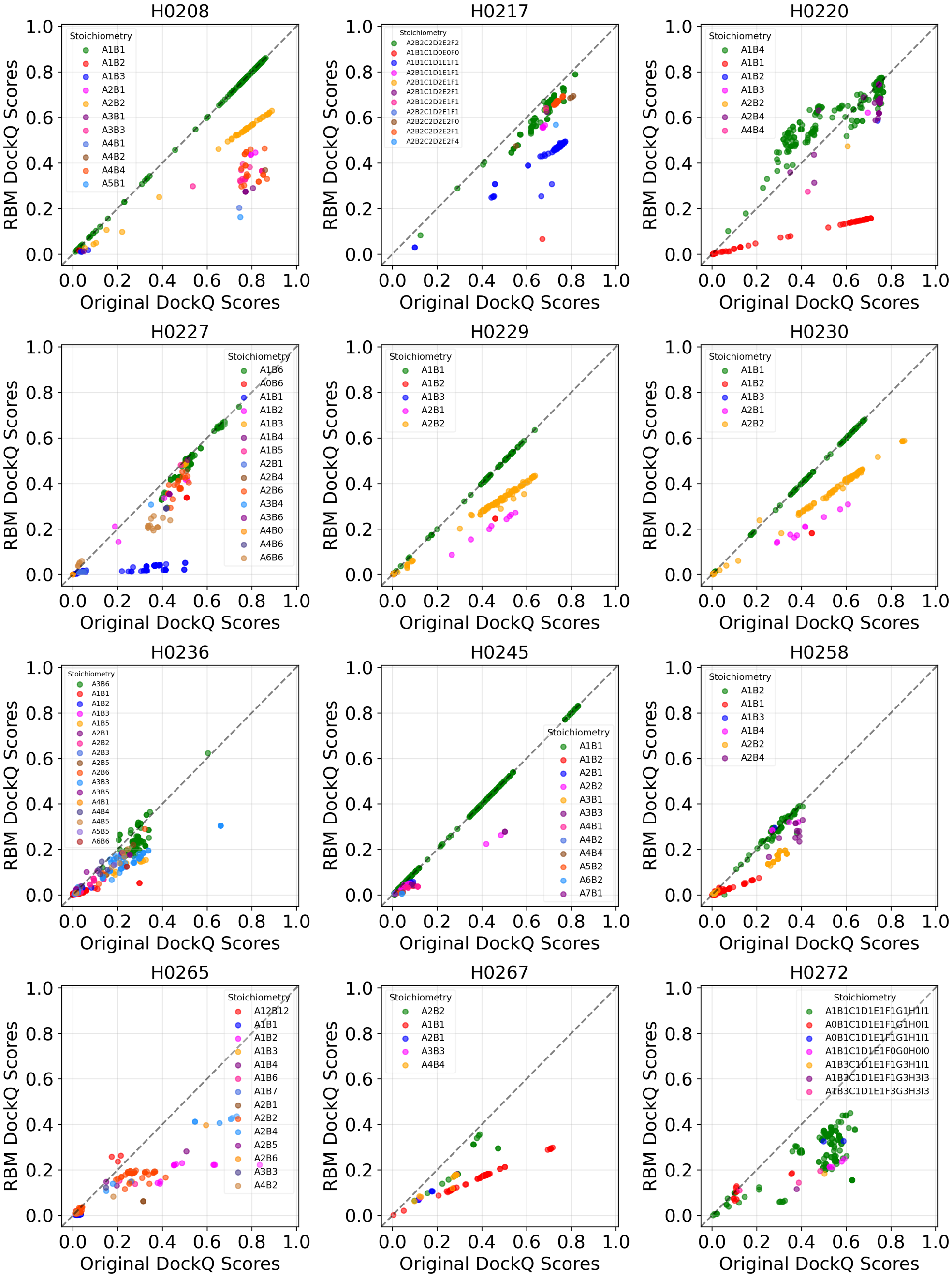

Figure 6: OST DockQ score vs. RBM DockQ score for Phase 0 hetero-oligomer targets.

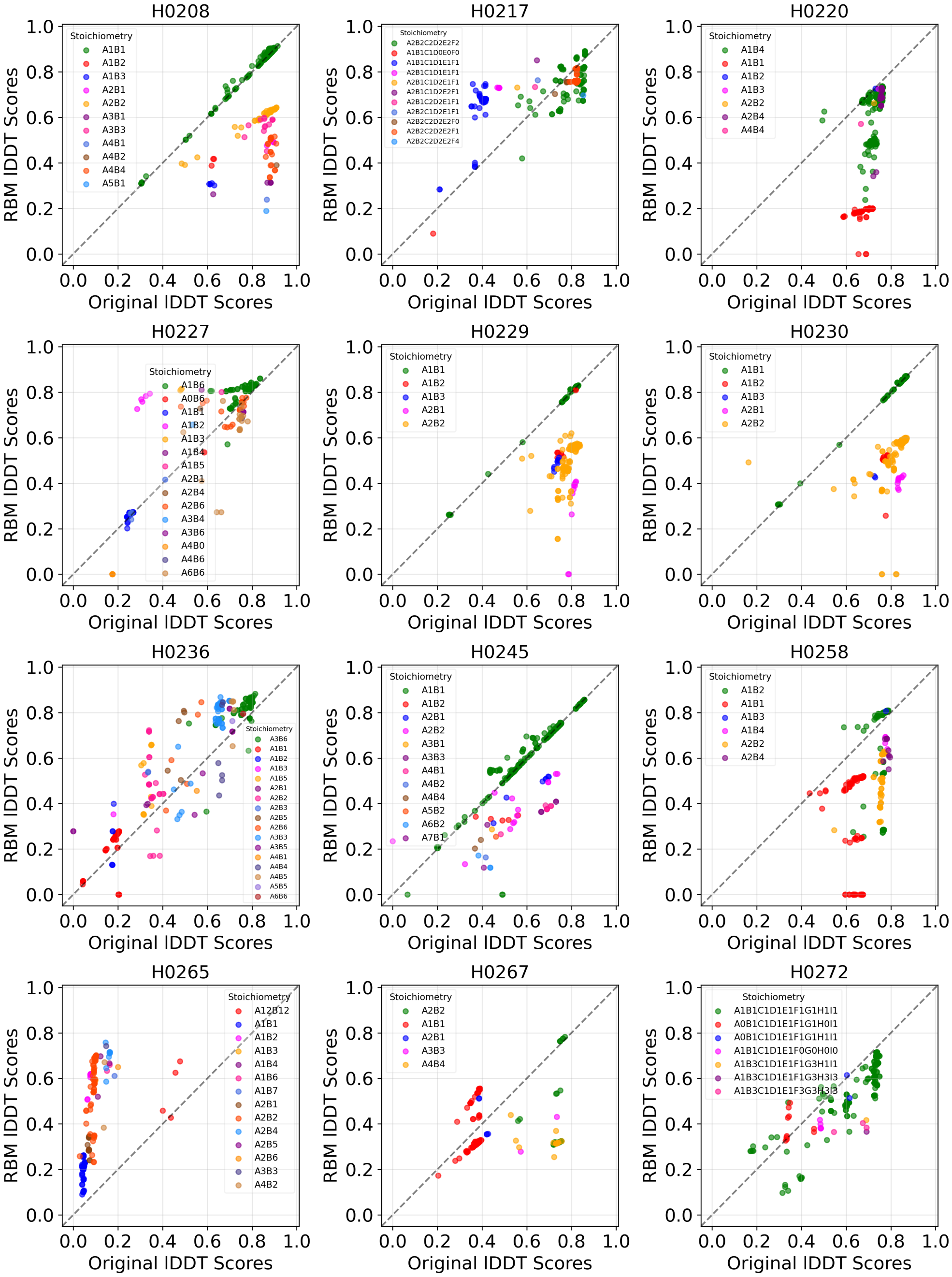

Figure 7: OST lDDT score vs. RBM lDDT score for Phase 0 hetero-oligomer targets.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims are achieved with our newly developed method.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper proposes a new scoring strategy and has no limitations at the moment. However, limitations between different versions of our methods are discussed.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not include theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We released the code, instructions, and example data, but the source data is not publicly available. However, they could be obtained by contacting casp@predictioncenter.org.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [N/A]

   Justification: This paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: This paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: This paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We conform with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Positive impacts are discussed, and the method we described should have no negative impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All software related to this work are open source and properly cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We have documents and GitHub repository.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Not applicable

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: Not applicable.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.