BMC Bioinformatics

SOFTWARE

Open Access

Reciprocal best matching: a new pipeline for scoring models with unknown stoichiometry in CASP experiments

Rongqing Yuan[1,2,3], Jing Zhang[1,2] and Qian Cong[1,2,4*]

*Correspondence:
Qian Cong
qian.cong@utsouthwestern.edu
[1]Eugene McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, TX, USA
[2]Department of Biophysics, University of Texas Southwestern Medical Center, Dallas, TX, USA
[3]Lyda Hill Department of Bioinformatics, University of Texas Southwestern Medical Center, Dallas, TX, USA
[4]Harold C. Simmons Comprehensive Cancer Center, University of Texas Southwestern Medical Center, Dallas, TX, USA

Abstract

Background 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, might introduce systematic biases when evaluating models with stoichiometries different from the target, sometimes favoring those with excess subunits and inflating scores for models with incorrect stoichiometries.

Results 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.

Conclusions Our method, RBM, could correct the systematic bias in the existing assessment protocol for protein complex structure prediction without stoichiometry information. We provide a standalone software implementation of our RBM pipeline to stimulate further method development in protein complex structure prediction and to support future CASP experiments.

Keywords CASP assessment, Protein complex structure prediction, Stoichiometry prediction

Background

The recent Critical Assessment of Structure Prediction Round XVI (CASP16) experiment shows that protein complex structure prediction remains an unsolved challenge [1, 2]. 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

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, 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 licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence 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 licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

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 [3–5], 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 Phase 1, stoichiometry information was unavailable for certain filamentous targets, and the predictors should predict the repetitive structural unit of the filaments. The need to predict stoichiometry in these challenges 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 [6–8]. Furthermore, predicted stoichiometry and protein complexes can assist the interpretation and validation of experimental structures. For example, it remains challenging to identify biological assemblies from crystal contacts [9, 10], and resolved crystal structures do not necessarily resemble the assembly under physiological conditions [11].

In CASP experiments, most complex assessment scores are calculated using a software package called OpenStructure (OST) [12]. These scores can be divided into two categories. Interface-focused metrics evaluate the quality of contacts between interacting chains (e.g., ICS, IPS, QS_best [13], and DockQ [14]), whereas global metrics (e.g., lDDT [15] and TM-score [16]) measure overall structural similarity between a prediction and the reference target. These scores have been widely used in previous CASP assessments [3, 13, 17, 18].

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, as CASP16 assessors, 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 interfaces (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 a model. Although no predictors deliberately exploited this caveat to achieve artificially higher scores, this feature of OST could result in a systematic bias.

In addition, we and previous CASP assessors [18] have noticed that the overall interface scores, such as ICS, IPS, and QSbest, 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, and the weight of each interface is thus close to the number of residues in this interface. 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.

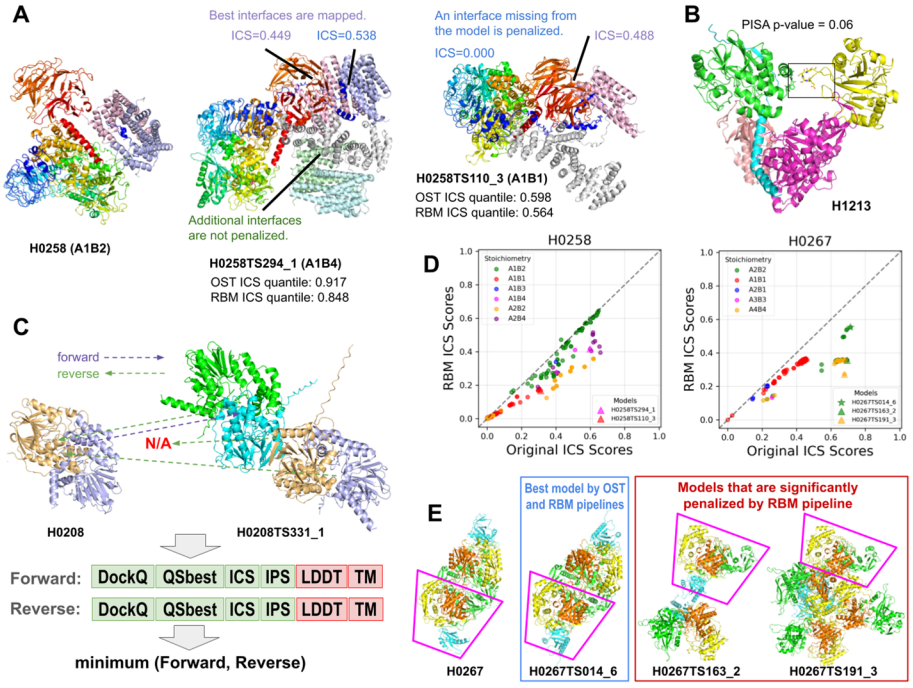

Fig. 1 **A** A systematic bias exists in the original CASP16 assessment method: predicting fewer subunits is penalized more than predicting more subunits. **B** Small interfaces can be biologically meaningful. **C** Graphic illustration of the RBM approach. Each interface in the target is mapped to the best interface in the model to compute a forward per-interface score (colored in green). Similarly, each interface in the model is mapped to the best interface in the target to compute the reverse per-interface score (colored in purple). These per-interface scores are weighted averaged to obtain the model-level forward and reverse scores, and the minimum of them is taken as the score for a model. **D** Example of ICS scores computed using the OST software and the RBM pipeline. Models with the correct stoichiometry are colored in green. The 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 (green and cyan are domains from chain **A**; orange and yellow are domains from chain **B**), the building block (in magenta trapezoid, which consists of one chain A and one chain **B**) is correctly predicted in all these models, but they are assembled incorrectly in the two models in the red box that are strongly penalized by our RBM pipeline

As the CASP16 assessors, we proposed to address these challenges by introducing the Reciprocal Best Matching (RBM) framework for evaluating protein complex prediction without known stoichiometry. Instead of directly comparing whole assemblies, RBM matches each interacting chain pair in the target to its best-matching pair in the model, and vice versa. Scores are computed in both directions, and the minimum is taken, so the method penalizes both missing interfaces and extra interfaces in the model, making it a stoichiometry-agnostic, unbiased alternative to the standard OST evaluation.

While the concept of RBM was introduced in our CASP16 assembly assessment paper, that work primarily focused on assessing predictors' performance and gaining new insights for the structure prediction community, with a very limited space for methodological exposition. In this study, we aimed to provide a comprehensive, reproducible, and methodologically detailed description of the RBM pipeline, including the conceptual motivation, algorithmic design, results visualization, and choices of key parameters. This work establishes RBM as a reproducible and extensible scoring framework intended for broader adoption beyond CASP16.

Implementation

The RBM pipeline takes as input the PDB structures of both the model and the target, as well as the JSON output files generated by OST. RBM extracts the definitions of contacting residues and the chemical groups of target and model chains (chem_groups and chem_mapping) from OST outputs, if provided. Otherwise, it computes contacting residues using the same criteria as OST; it assigns the chemical groups between target and model chains based on sequence identity, considering chains with identity above 99% as the same protomer and assigns them to the same group.

RBM operates bidirectionally: all interfaces in the target are matched to their best corresponding interfaces in the model (forward matching), and conversely, all interfaces in the model are matched to their best corresponding interfaces in the target (reverse matching). The resulting matched interfaces are then used to compute a range of evaluation scores, including, but not limited to, ICS, IPS, QSbest, lDDT, TMscore, and DockQ. When a single target interface matched multiple candidate model interfaces, we retained the maximum value for each score. Any interface that appears exclusively in either the model or the target is penalized by assigning a score of 0 (Fig. 1C). It is worth noting that RBM does not assume or enforce a one-to-one chain mapping; rather, multiple chain pairs in the model may map to the same chain pair in the target. Abandoning one-to-one chain mapping is a key distinction between RBM and existing approaches such as QS_global, and is essential for evaluating predictions with unknown or incorrect stoichiometry.

In our implementation, ICS, IPS, and QSbest are computed for each pair of matched interfaces based on the following formulas using the contacts and interface residues (residues that participate in inter-chain contacts): (1) $ICS = 2*\frac{precision(contacts)*recall(contacts)}{precision(contacts)+recall(contacts)}$; (2) $IPS = \frac{|M_i \cap T_i|}{|M_i \cup T_i|}$, where $M_i$and T_i are the set of interface residues in the model and the target, respectively; (3) $QSbest = \frac{|M_{cnt} \cap T_{cnt}|}{max(M_{cnt}, T_{cnt})}$, where M_{cnt} and T_{cnt} are the sets of contacts in the model and the target, respectively.

We tested multiple distance cutoffs for defining inter-chain contacts and compared the resulting scores with those computed by OST. We selected the cutoff that minimized the discrepancy between our scores and OST's. For ICS and IPS, two residues from different chains are considered in contact if the minimum heavy-atom distance between them is ≤5 Å, and both residues are considered to be interface residues. For QS_best, we used a 10 Å cutoff for heavy-atom distance to define contacts. lDDT, TM-score, and DockQ values for each pair of matched chains between the target and the model are calculated using external software.

RBM pipeline evaluates one pair of interacting chains against its matched chain pair at a time. Scores for all interacting chains in the target are subsequently weight-averaged to yield a "forward" score. Meanwhile, the scores for all interacting chains 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 $log_{10}Nave$ as the weight, where $Nave$ is the average number of interface residues (criterion: heavy-atom distance to another chain < 5 Å) for an interface (in forward scores, the interface is from the target and in reverse scores, the interface is from the model).

Compared to OST's interface-weighting strategy, our approach upweights smaller interfaces, thereby rewarding successful predictions of these more challenging interactions.

Notably, 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[2]. 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.

Results

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 widely used in CASP complex assessments. As shown in Fig. 1D, ICS scores generated by RBM remain highly correlated with those computed by OST. At the same time, models with incorrect stoichiometry tend to receive lower scores under the RBM pipeline. Notably, RBM and OST scores differ the most for models that predict more subunits than the target. This is related to the fact that OST might be biased towards such predictions (overprediction of subunits) by finding the best subcomplex within a model to maximize the similarity to the target, whereas RBM will penalize any incorrectly predicted interfaces. In our assessment of CASP16, we initially used a version of the 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 (Figs. 1E and Fig. S1). 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 and the reverse scores. This method yields a distribution of scores shown in Fig. S3 and penalizes models with incorrect stoichiometry, whether they over- or under-predict subunit numbers, consistent with the goal of 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 Figs. S1–S3, we provide the distributions of ICS scores under 3 different strategies for all Phase 0 heteroligomer targets, excluding antibody-antigen targets [2]. Because no predictors in CASP16 attempt to exploit the potential biases with OST or the original version (averaging forward and reverse scores) of RBM to obtain high scores, different strategies do not remarkably affect the ranking of CASP16 groups. Nevertheless, we believe our favored strategy, i.e., taking the minimum of the forward and reverse scores,

is the most robust. We therefore considered this strategy the default for the RBM pipeline and used it in our CASP16 evaluation paper. In addition, when combining the scores of different interfaces, RBM allows the user to choose either $Nave$ or $log_{10}Nave$ as the per-interface weight, where $Nave$ is the average number of interface residues between the two interacting chains. By default, we use $log_{10}Nave$ to upweight smaller interfaces. For comparison, we also report ICS scores computed using $Nave$ as the weight (Fig. S4). The resulting score distributions and rankings of CASP16 groups are similar, indicating that emphasizing smaller interfaces does not lead to major changes in overall evaluation outcomes.

To study the difference between RBM scores and OST scores, we grouped models into three categories – correct stoichiometry, more subunits, and less subunits – and examined the distributions of either score, as well as the differences between them (Fig. S5). Overall, the two scoring routines show the largest difference in models that contain more subunits than the target. Nevertheless, the potential bias towards models with more subunits would not have remarkably affected the CASP16 rankings because the Phase 0 experiment was newly introduced, and predictors had not yet optimized their strategies to win this category. We have computed the ranking over Phase 0 targets (excluding antibody-antigen targets) using the OST scores and the RBM scores, respectively. While the top three groups remain the same in both rankings, the ranks of many other groups have changed slightly.

The differences in the sum z-score over phase 0 targets based on the OST scores and the RBM scores for each group are shown in Fig. S6. Groups with higher z-scores by RBM strategy tend to have correct stoichiometry predictions for a larger number of targets. For example, TS322 (XGroup) increased the most in RBM-based z-scores relative to OST-based z-scores, and has also predicted the correct stoichiometry for the largest number of targets [2]. Despite the lack of remarkable changes in the ranking, the groups showing better performance in stoichiometry prediction are rewarded.

We also applied the RBM routine to other widely used CASP scores, such as DockQ (Fig. S7) and lDDT (Fig. S8), and compared the values with scores computed by OST. For DockQ, the RBM-adjusted scores remain highly correlated with the OST scores. lDDT scores computed by RBM and OST showed a weaker correlation between the RBM and OST; this is related to the fact that the RBM pipeline computed lDDT scores in a pairwise fashion, just like it did for other interface scores. This pairwise-averaging approach may not be optimal for lDDT and TM-score, which are intended to evaluate the global structure of the entire target. An alternative solution could be to perform a "bidirectional alignment" between the entire target and model structures, rather than averaging over pairwise scores.

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 [19–22] 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/conglab2020/RBM.

Abbreviations

CASP Critical assessment of structure prediction
OST OpenStructure

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-026-06439-7.

Additional file 1.

Acknowledgements

We thank all CASP organizers, in particular, Andriy Kryshtafovych and John Moult, for their guidance during our CASP16 assessment. 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 in our CASP16 assessment. We also acknowledge computing resource support from NSF ACCESS award MED240004, TACC award MCB23014, and NIAIR award NAIRR240466.

Author contributions

RY and JZ developed the method with supervision from QC. RY and QC developed the software. RY, JZ, and QC analyzed the results. RY and QC prepared the figures and drafted the manuscript. RY, JZ, and QC edited the manuscript.

Funding

QC is supported by the Endowed Scholars Program in UTSW and 1R35GM160468-01 from NIGMS. JZ is supported by NIAID K99AI180984-01A1.

Data availability

The pdb structure data from CASP experiments is available for download at https://predictioncenter.org/casp16/results.cgi?view=targets&tr_type=multimer. To facilitate reproducibility, we have curated thedatasets from CASP and deposited them on Zenodo at https://doi.org/10.5281/zenodo.18369915. The code and example data used in this paper are uploaded to https://github.com/conglab2020/RBM.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Availability and Requirements

Project name: Reciprocal Best Matching. Project home page: https://github.com/conglab2020/RBM.Operating system(s): UNIX-based systems. Programming language: Python, Shell. Other requirements:Python 3.9 or higher; DockQ, IDDT, TM-align (or US-align). License: MIT. Any restrictions to use by nonacademics:None.

Received: 7 November 2025 / Accepted: 30 March 2026

Published online: 17 April 2026

References

1. Yuan R, et al. CASP16 Protein monomer structure prediction assessment. Proteins Struct Funct Bioinforma. 2026;94:86–105.
2. Zhang J, et al. Assessment of protein complex predictions in CASP16: are we making progress? Proteins Struct Funct Bioinforma. 2026;94:106–30.
3. Guzenko D, Lafita A, Monastyrskyy B, Kryshtafovych A, Duarte JM. Assessment of protein assembly prediction in CASP13. Proteins Struct Funct Bioinforma. 2019;87:1190–9.
4. Kryshtafovych A, et al. Updates to the CASP Infrastructure in 2024. Proteins Struct Funct Bioinforma. 2026;94:15–24.
5. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Progress and bottlenecks for deep learning in computational structure biology: CASP Round XVI. Proteins Struct Funct Bioinforma. 2026;94:5–14.
6. Schmidt C, Lenz C, Grote M, Lührmann R, Urlaub H. Determination of protein stoichiometry within protein complexes using absolute quantification and multiple reaction monitoring. Anal Chem. 2010;82:2784–96.
7. Zilkenat S, Grin I, Wagner S. Stoichiometry determination of macromolecular membrane protein complexes. Biol Chem. 2017;398:155–64.

8. Degliesposti G. Probing protein complexes composition, stoichiometry, and interactions by peptide-based mass spectrometry. In Advanced Technologies for Protein Complex Production and Characterization: Volume II (eds Vega, M. C. & Fernández, F. J.) 41–57Springer Nature Switzerland, Cham, (2024). https://doi.org/10.1007/978-3-031-52193-5_4
9. Dey S, Ritchie DW, Levy ED. PDB-wide identification of biological assemblies from conserved quaternary structure geometry. Nat Methods. 2018;15:67–72.
10. Dey S, Levy ED. PDB-wide identification of physiological hetero-oligomeric assemblies based on conserved quaternary structure geometry. Structure. 2021;29:1303–e13113.
11. Krissinel E, Henrick K. Inference of macromolecular assemblies from Crystalline State. J Mol Biol. 2007;372:774–97.
12. Biasini M, et al. OpenStructure: an integrated software framework for computational structural biology. Acta Crystallogr D Biol Crystallogr. 2013;69:701–9.
13. Lafita A, et al. Assessment of protein assembly prediction in CASP12. Proteins Struct Funct Bioinforma. 2018;86:247–56.
14. Basu S, Wallner B, DockQ. A Quality measure for protein-protein docking models. PLoS ONE. 2016;11:e0161879.
15. Mariani V, Biasini M, Barbato A, Schwede T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics. 2013;29:2722–8.
16. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005;33:2302–9.
17. Ozden B, Kryshtafovych A, Karaca E. Assessment of the CASP14 assembly predictions. Proteins Struct Funct Bioinforma. 2021;89:1787–99.
18. Ozden B, Kryshtafovych A, Karaca E. The impact of AI-based modeling on the accuracy of protein assembly prediction: Insights from CASP15. Proteins Struct Funct Bioinforma. 2023;91:1636–57.
19. Elofsson A. AlphaFold3 at CASP16. Proteins Struct Funct Bioinforma. 2026;94:154–66.
20. Liu J, Neupane P, Cheng J. Accurate stoichiometry prediction of protein complexes by integrating AlphaFold3 and template information. 2025.01.12.632663 Preprint at https://doi.org/10.1101/2025.01.12.632663 (2025).
21. Liu J, Neupane P, Cheng J. Improving AlphaFold2- and AlphaFold3-Based protein complex structure prediction With MULTICOM4 in CASP16. Proteins Struct Funct Bioinforma. 2026;94:131–41.
22. Wang W, Luo Y, Peng Z, Yang J. Accurate biomolecular structure prediction in CASP16 with optimized inputs to state-of-the-art predictors. Proteins Struct Funct Bioinforma. 2026;94:142–53.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.