═══REVIEW COMMONS MANUSCRIPT═══

IMPORTANT:

- Manuscripts submitted to Review Commons are peer reviewed in a journal-agnostic way.
- Upon transfer of the peer reviewed preprint to a journal, the referee reports will be available in full to the handling editor.
- The identity of the referees will NOT be communicated to the authors unless the reviewers choose to sign their report.
- The identity of the referee will be confidentially disclosed to any affiliate journals to which the manuscript is transferred.

GUIDELINES:

- For reviewers: https://www.reviewcommons.org/reviewers
- For authors: https://www.reviewcommons.org/authors

CONTACT:

The Review Commons office can be contacted directly at: office@reviewcommons.org

Reconstructing Biological Molecules with Help from Video Gamers

[1]Andreas C. Petrides, *[2]Robbie P. Joosten, Foldit Players, [1]Firas Khatib, *[3]Scott Horowitz

Correspondence: scott.horowitz@du.edu, r.joosten@nki.nl

[1]Department of Computer and Information Science, University of Massachusetts Dartmouth, Dartmouth, Massachusetts, United States of America

[2]Oncode Institute and Division of Biochemistry, Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands

[3]Department of Chemistry & Biochemistry and The Knoebel Institute for Healthy Aging, University of Denver, Denver, Colorado, United States of America.

Abstract
Foldit is a citizen science video game in which players tackle a variety of complex biochemistry puzzles. Here, we describe a new series of puzzles in which Foldit players improve the accuracy of the public repository of experimental protein structure models, the Protein Data Bank (PDB). These efforts highlight the continued need for the engagement of the lay population in science.

Introduction
Our shared knowledge of experimental biomolecular structures is stored in the Protein Data Bank (PDB) (*1*). The accuracy of the PDB is important for multiple reasons. For example, the recent successes in deep learning approaches to predicting protein structure (*2*) rely on an accurate PDB for continued development. Similarly, the PDB serves as a resource for biologists and biochemists to develop hypotheses to test in their everyday experiments. Mistakes in structure models can cause scientists to base their hypotheses on incorrect data, which causes delays in scientific progress.

The PDB contains many solvable errors (*3*). The type of data used in crystallography and cryo-EM, maps of electron density or electron potential, respectively, have limitations as they are the result of an indirect experiment with experimental error and limited resolution. By showing where the electrons are, scientists can fit atoms to these maps to discover the atomic structure of the macromolecule in the experiment. This requires a combined approach of visual fitting of these maps as well as computational tools to aid in complying with known rules of chemistry and physics in how molecules are put together. Limitations include: low resolution of the data, parts of the maps where artifacts or missing data can cause blurriness in the maps, assumptions in computational models used to process the primary experimental data, among other causes of low data quality. As a result, the scientists interpreting these data can make mistakes. High attention to detail and the use of verification tools can help prevent some mistakes (*4*), but the sheer quantity of data makes it likely that human error will still persist into published and deposited structure models. Furthermore, the PDB does not require peer review for deposited

models, and as a result, many entries that contain errors ranging from inconsequential to egregious exist within the PDB.

For this reason, the PDB-REDO project was begun in 2006 (*5*). The mission of PDB-REDO is to perform automated re-refinement and rebuilding of the crystallographic structure models in the PDB to improve the accuracy of the PDB and remove model errors in the process. This venture has been highly successful, with the PDB-REDO databank now containing over 170k entries, many of which have an improved fit to the experimental data and more probable structural features (*6*).

However, is there still a need for improvement of structure models beyond the automated re-refinement by PDB-REDO? Yes, the procedure in PDB-REDO looks at many structural aspects (*7*) but is limited to model issues that can be handled robustly with a very low risk of making the model worse. Indeed, a search of the PDB for its lowest quality structure models shows that the improvements by PDB-REDO in these cases can be modest at best. They still appear to be relatively low quality after PDB-REDO refinement. As a result, new approaches are needed.

One possible approach is to enlist the aid of humans to improve the PDB via the players of the biochemistry video game Foldit. Foldit is a citizen science game in which players work on a variety of complex biochemistry puzzles, collaborating and competing to create the best possible structure model (*8*). Model quality within the game is judged by the Rosetta force field (*9*), combined with other elements, such as the fit to an underlying map from experimental data (*10*). Previous competitions have shown that Foldit players can solve both crystal structures and cryo-EM structures with higher accuracy than scientists or computational algorithms (*10, 11*). Therefore, it became an obvious question to ask whether Foldit players could also improve the structures already within the PDB?

Results

To test this possibility, we created a new series of Foldit puzzles, termed *Reconstruction puzzles*. In these puzzles, Foldit players were given protein structure models and the underlying density maps and tasked with improving the structures. In this report, we discuss the results after the first 58 of these puzzles.

After the puzzles, the players' solutions were clustered and then analyzed by PDB-REDO for comparative analysis to see if using the Foldit structure models as a starting point is an improvement over using the structure models from the PDB itself. In 39 of the 58 cases, the overall structure quality was improved by using the Foldit structure. In general, the Foldit player structures had considerably improved chemical and physical properties in the vast majority of puzzles by multiple metrics, and in many cases also improved the fit to the experimental data (Fig 1 and Supplemental Table 1). As seen by Molprobity ClashScore (*12*) (Fig 1C), the atomic clashes of the structure models were especially improved. Two examples of before and after the Foldit use can be seen in Fig 2. In some cases, such as in Fig 2A, changes to the structure can be observed at many points. In many of the cases though, such as in Fig 2B, the changes are localized, and often in regions of poor electron density such as loops. However, these regions

can be important biologically. For example, in the case of 2fxt, yeast mitochondrial import inner membrane translocase subunit TIM44, we observe that Gly388 has changed position by 3.2 Å due to a so-called register shift. This glycine is fully conserved over 1522 sequences in HSSP (*13*) and in the more distant human ortholog (Uniprot AC: O43615) this glycine is conserved as well. Moreover, AlphaMissense (*14*) suggests that mutations to this residue are very likely to be pathogenic (minimal score 0.922) indicating that this a key residue in the protein.

With the success of the *Reconstruction puzzle* series, we have created a new pipeline in which data from Foldit puzzles can be automatically analyzed and evaluated by PDB-REDO for inclusion in the PDB-REDO database. In these cases, when the Foldit player has agreed to have their Foldit name attached to the entry, the PDB-REDO webpage will give credit to the Foldit player whose structure model is now available for public download.

How did the Foldit players improve on these structures? To get an idea, we asked several Foldit players whose structure models were chosen as the best overall structure. The approaches taken by different players was quite variable. In most cases, the strategy involved a combination of hand-fitting as well as automated tools and scripts. However, in certain cases, the top solutions used only scripts and automated tools, suggesting that considerable improvement in automated fitting of structural maps can be accomplished (See Supplemental Information).

Conclusions

With the success of the Foldit players in improving experimental structure models, the *Reconstruction Puzzle* series within Foldit continues, and will continue to provide improvements to known structures. At present, these puzzles are only for crystallographic structure models, but with the increase in number of cryo-EM structure models, the same problems persist there as well, and can also use the intervention of Foldit players.

With the focus on new Artificial Intelligence tools and how they can improve fields including structural biology, cases like this are here to remind us that the greatest untapped potential for science is that of humanity.

Methods

Reconstruction puzzles were prepared by performing 5 rounds of refinement in Phenix (*15*) using the original PDB entry with all ligands and waters removed and including simulated annealing. This new model was given to the players with a Feature Enhanced Map (*16*) for reconstruction.

A set of post processing steps occurred following each puzzle's completion:

Using retrieved player-generated solution files, a Python script orchestrated the clustering of PDB files based on their structural similarity, employing the Cα Root Mean Square Deviation (CA-RMSD) metric. Iteratively, the script varies the threshold for CA-RMSD clustering, starting from 1 Å and decreasing until the total number distinct clusters is greater than 100. This iterative

clustering process is crucial since player-generated solutions contain a diverse set of structural configurations. This set of steps filtered the structures down into a concise set of models.

Following clustering, the solutions were clustered and ranked based on their Foldit score, with the top 100 clustered solutions moving to the next step. The Python script then submitted the top 100 clustered solutions by score to the API of the PDB-REDO web service (*17*), facilitating the submission of these clustered solutions for further refinement and analysis.

The PDB-REDO procedure used here entailed optimizing the weight between the experimental data (as retrieved from the PDB) and restraints for covalent geometry and atomic B-factors to maximize the fit with the experimental data whilst maintaining normal geometry. Combinations of different types of additional restraints were used while refining the models with REFMAC (*18*): homology-based hydrogen bond restraints (*6*) were applied with 8 solution sets, general hydrogen bond restraints with 18 sets, local non-crystallographic symmetry restraints with 26 sets and jelly-body restraints with 19 sets (*19*). All the model rebuilding steps in PDB-REDO were excluded to stay close to the original solution.

The Python script awaited confirmation of completion for each solution from the PDB-REDO API. Upon completion, the finalized protein structures and associated metrics were retrieved. Deposited structures were written into new PDB files. Corresponding metrics were organized into comprehensive data dictionaries and graphs, both of which facilitate further analysis.

Acknowledgements: We would like to thank J. Flatten for his assistance maintaining the Foldit platform and the RHPC facility of the Netherlands Cancer Institute for maintaining PDB-REDO's compute infrastructure. Funding was provided by NSF 2051305 to SH, NSF 2051282 to FK, Horizon 2020 research and innovation program grant No. 871037 (iNEXT-Discovery) to RPJ and by an institutional grant of the Dutch Cancer Society and of the Dutch Ministry of Health, Welfare and Sport.

The authors declare that they have no conflicts of interest.

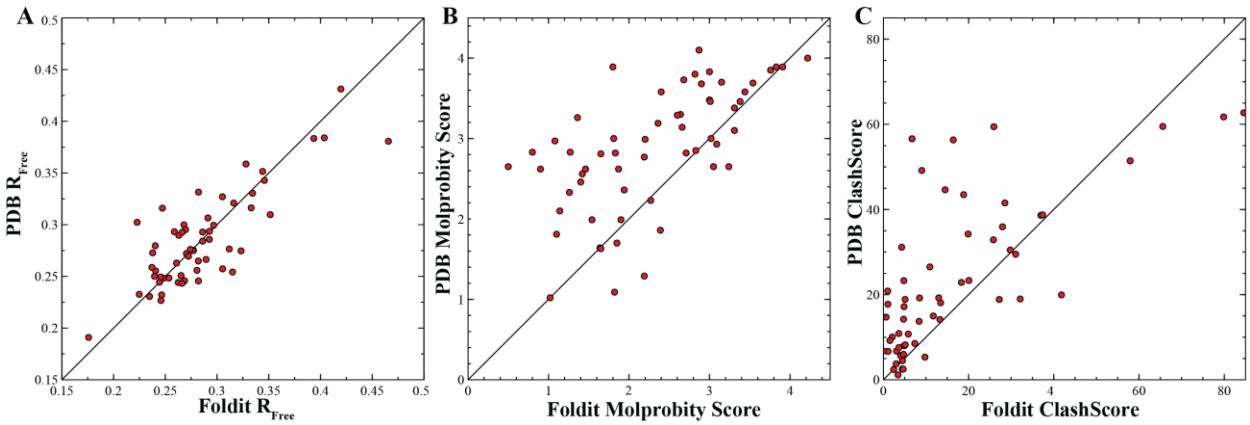

Figure 1: Comparison of selected statistics for PDB-REDO output using original PDB entry (y-axis) or Foldit-refined entry (x-axis) as input. Each point is a separate PDB entry. A. R_{Free}, B.

Molprobity Score, and C. Molprobity ClashScore for each entry. As lower scores are better, above the diagonal line are cases in which the Foldit model improved over the original PDB entry as the starting point for PDB-REDO.

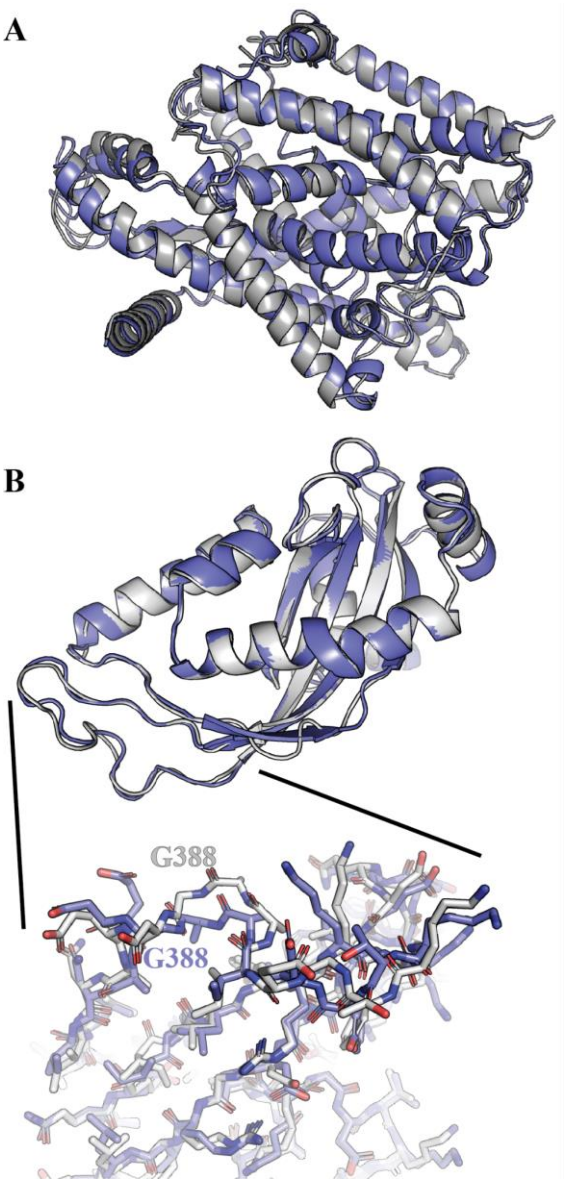

Figure 2. Structure of 2yxq (A) and 2fxt (B). Foldit structure used as starting point for PDB-REDO in blue, PDB-REDO using original PDB as starting structure in gray.

1. H. Berman, K. Henrick, H. Nakamura, Announcing the worldwide Protein Data Bank. *Nature Structural Biology* **10**, 980-980 (2003).
2. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).
3. R. P. Joosten, T. Womack, G. Vriend, G. Bricogne, Re-refinement from deposited X-ray data can deliver improved models for most PDB entries. *Acta Crystallogr D* **65**, 176-185 (2009).

4. R. J. Read *et al.*, A new generation of crystallographic validation tools for the protein data bank. *Structure* **19**, 1395-1412 (2011).
5. R. P. Joosten, G. Vriend, PDB Improvement Starts with Data Deposition. *Science* **317**, 195-196 (2007).
6. B. van Beusekom *et al.*, Homology-based hydrogen bond information improves crystallographic structures in the PDB. *Protein Sci* **27**, 798-808 (2018).
7. R. P. Joosten, K. Joosten, G. N. Murshudov, A. Perrakis, PDB-REDO: constructive validation, more than just looking for errors. *Acta Crystallogr D* **68**, 484-496 (2012).
8. S. Cooper *et al.*, Predicting protein structures with a multiplayer online game. *Nature* **466**, 756-760 (2010).
9. J. K. Leman *et al.*, Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat Methods* **17**, 665-680 (2020).
10. S. Horowitz *et al.*, Determining crystal structures through crowdsourcing and coursework. *Nature Communications* **7**, (2016).
11. F. Khatib *et al.*, Building de novo cryo-electron microscopy structures collaboratively with citizen scientists. *Plos Biol* **17**, (2019).
12. V. B. Chen *et al.*, MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* **66**, 12-21 (2010).
13. R. P. Joosten *et al.*, A series of PDB related databases for everyday needs. *Nucleic Acids Res* **39**, D411-419 (2011).
14. J. Cheng *et al.*, Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, 1303-+ (2023).
15. D. Liebschner *et al.*, Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in. *Acta Crystallogr D* **75**, 861-877 (2019).
16. P. V. Afonine *et al.*, FEM: feature-enhanced map. *Acta Crystallogr D* **71**, 646-666 (2015).
17. R. P. Joosten, F. Long, G. N. Murshudov, A. Perrakis, The PDB_REDO server for macromolecular structure model optimization. *Iucrj* **1**, 213-220 (2014).
18. G. N. Murshudov *et al.*, REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr D Biol Crystallogr* **67**, 355-367 (2011).
19. A. A. Vagin *et al.*, REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use. *Acta Crystallogr D* **60**, 2184-2195 (2004).