

The Topological Properties of the Protein Universe

Corresponding Author: Professor Michael Stumpf

This manuscript has been previously reviewed at another journal. This document only contains information relating to versions considered at Nature Communications.

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

Version 0:

Reviewer comments:

Reviewer #2

(Remarks to the Author)

I'll keep my comments brief as I'm rereviewing it based on a previous review the authors commented on. I think this paper should be published after some minor revisions/edits.

This paper shows that topology is ready to contribute in interesting ways. The findings are broadly supported (see a few notes below).

Notes:

I still find Figure 1 fails as a helpful figure. The details are too obscured by the large bubbles. The color coding fails, etc. I appreciate the challenge of showing a few hundred million datapoints on a single plot, but I think the circle packing obscures things. I would suggest something much easier to make it easier to find interesting points in the space of topological richness.

The new study is nice (Figure 2), although the conclusion (eukaryota is different from archae and bacteria) doesn't strike me as overly surprisingly. Is there some aspect of this result that might be more notable? Is there some interpretation of your findings that would suggest proteins useful for multicellular life?

Notes:

Supplemental Figure 11 continues to evaluate something different than 'statistical significance' (see line 276). As far as I can tell there hasn't been a statistical test computed on these, so the language needs to change or you need a statistical test. To be clear, I appreciate having the error bars which I find more helpful, but the matter of precision remains.

There were also a number of minor comments the authors didn't reply to from the previous review. I'm repeating those I still think are relevant (I did see some edits in response to these points.)

The topological and functional relevance of voids is fairly intuitive. Why are 1d loops relevant? (I realize you show they help identify CATH domains, but would that have been expected?)

The proposed topology ignores all the existing recognized structures in the proteins, including things like alpha-helices and beta-strands and beta-sheets. While I understand the idea of ignoring this structure when trying to apply existing topological tools, it seems like ignoring this structure entirely is perhaps too crude a path. Is there no way to use adapt the topological tools to merge these? Alternatively, perhaps add some supplemental discussion on why this is unlikely to be needed in terms of the results on the CATH domains?

I was not able to figure out what Figure 4B was showing. This plot needs better labels or more explanation or some other context. The legend seems to make it seem like there should be an interpretation of error bars in this figure, but I wasn't able to identify them.

For the error bands on Figure 4C, why not show density plots like the rest of the paper? One could also compute a KS

statistical test of different distributions to make a more precise comparison.

(Remarks on code availability)

I didn't try running the code. The code appears to be helpful documented including things like README and

Version 1:

Reviewer comments:

Reviewer #2

(Remarks to the Author)

Thank you for the detailed reply. This remains a good paper and the revisions make it even better. I remain skeptical of the utility of the main figure, but I view that as an editorial call and you (the authors) are obviously in favor. I see no reason that would prohibit publishing it.

(Remarks on code availability)

See previous review.

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/>

Second review round

We thank the referee for their helpful and thoughtful comments, which significantly improved the manuscript; crucially the analysis suggested by the reviewer has prompted us to compare and contrast topological features of proteins in unicellular and multicellular eukaryotes.

Below is a point by point response. Changes in the manuscript are highlighted in blue.

1. I still find Figure 1 fails as a helpful figure. The details are too obscured by the large bubbles. The color coding fails, etc. I appreciate the challenge of showing a few hundred million datapoints on a single plot, but I think the circle packing obscures things. I would suggest something much easier to make it easier to find interesting points in the space of topological richness.

We thank the reviewer for their thoughtful comment. In response, we explored several alternatives to circle packing, such as ellipse packing and tessellation as an area proportional Voronoi diagram. Unfortunately, none were more effective at highlighting the differences between domains or simplifying the figure. To enhance clarity, we added quantifications to the figure in the form of topological richness values for each of the highlighted species, for kingdoms and phyla, and each domain of life, aiming to make the figure more effective in quantifying global topological information.

We recognise that this adjustment alone may not fully address the challenge of presenting the fascinating features of the protein universe. To better incorporate the reviewer's suggestion and provide simpler summaries, we added panel A to Figure 2, and Table 1. The new panel contains a density plot for each domain of life showing topological richness aggregated from calculations on each individual high quality protein. This addition is intended to complement the existing figure by offering a straightforward summary of the results. We also provide a short paragraph explaining the results (lines 149-154).

Despite these changes, we believe that Figure 1 remains an important part of the paper, offering a graphical and intuitive overview of the work. Its primary purpose is to illustrate the scope of our study rather than to focus on specific results. For this reason, we decided to retain Figure 1 with slight modifications, ensuring it continues to add value to the manuscript.

2. The new study is nice (Figure 2), although the conclusion (eukaryota is different from archaea and bacteria) doesn't strike me as overly surprisingly. Is there some aspect of this result that might be more notable? Is there some interpretation of your findings that would suggest proteins useful for multicellular life?

We thank the reviewer for this insightful comment. In response, we have added brief paragraphs (lines 155-157, 165-178, 184-187) discussing the sharp peaks observed in the eukaryotic distributions, which contrast with the general uniformity seen in archaea and bacteria. We speculate that these peaks may indicate a specific level of protein complexity that facilitates the intricate regulatory mechanisms characteristic of multicellular lifeforms. To explore this further, the eukaryotic data was divided into two groupings for reanalysis, where a group for multicellular organisms was approximated by taking all proteins from Metazoan and Embryophyta (informally, animals and land plants), while the other group was the remaining eukaryotic proteins. The sharp peaks were found to be confined to the first grouping approximating multicellular organisms (see Suppl. Fig 17). We believe that this is an important result; we thank the reviewer for suggesting an analysis which pointed us in the right direction.

3. Supplemental Figure 11 continues to evaluate something different than 'statistical significance' (see line 276). As far as I can tell there hasn't been a statistical test computed on these, so the language needs to change or you need a statistical test. To be clear, I appreciate having the error bars which I find more helpful, but the matter of precision remains.

We performed a Mann–Whitney U test for the void volume values displayed in suppl fig 11, to show they are of different distributions, see lines 308-313..

4. The topological and functional relevance of voids is fairly intuitive. Why are **1d loops relevant?** (I realize you show they help identify CATH domains, but would that have been expected?)

In general, one-dimensional (persistent) homology is known to capture qualitative spatial features of a shape or structure. Given this, it is reasonable to expect meaningful insights in our case. Specifically, for proteins and spatial curves, previous work by some of the authors and collaborators has shown that 1D loops can subtly capture geometric substructures, including entanglement and other non-trivial spatial features. Our findings align with this perspective: we observe that 1D loops are intricately folded and interwoven within folding domains, making it unlikely for them to traverse distinct domains. In the simplest case, such

domains may be connected only by a single thin and flexible bond. We have added a paragraph discussing this in further detail (lines 241-249).

5. The proposed topology ignores all the existing recognized structures in the proteins, including things like alpha-helices and beta-strands and beta-sheets. While I understand the idea of ignoring this structure when trying to apply existing topological tools, it seems like ignoring this structure entirely is perhaps too crude a path. Is there no way to adapt the topological tools to merge these? Alternatively, perhaps add some supplemental discussion on why this is unlikely to be needed in terms of the results on the CATH domains?

Secondary structure is both fundamental and useful, but it cannot be inferred perfectly from sequence alone. Instead, it is predicted from sequence data and only becomes a standardised annotation when a known 3D structure is available. Since predicted structures are not always fully reliable, precisely determining the presence or absence of secondary structure is non-trivial. That said, this remains an interesting avenue for future exploration. We also note that in more restricted datasets, other authors have applied a persistent homology (PH) approach incorporating secondary structure or even atomistic details of protein chains. We have added a paragraph in the conclusions (lines 384-389) discussing this in further detail. In our case, the core reason for not incorporating secondary structure is to work directly with the fundamental raw data, allowing TDA to capture structural features without additional annotations influencing the analysis.

6. I was not able to figure out what Figure 4B was showing. This plot needs better labels or more explanation or some other context. The legend seems to make it seem like there should be an interpretation of error bars in this figure, but I wasn't able to identify them.

We thank the reviewer for this comment. The caption text of Figure 4B has now been updated. We trust that the new caption provides a clearer description of the figure.

7. For the error bands on Figure 4C, why not show density plots like the rest of the paper? One could also compute a KS statistical test of different distributions to make a more precise comparison.

We thank the referee for the comment. The other density plots referenced in the main text are the violin plots shown in Figure 5A and 5B, which also display TIFs, but with only three comparisons in each of the two subfigures. We think that in this case,

20 comparisons in Figure 4C might appear too cramped if presented as violin plots. We have also performed statistical tests to make a more precise comparison, and added a paragraph (lines 314-332) discussing these results, as well as Suppl.Fig.16 and Suppl.Table3).