Qeios

Peer Review

Review of: "Mapping Cryptic Phosphorylation Sites in the Human Proteome"

Benjamin C. Orsburn[1]

1. Organ Pathobiology and Therapeutics Institute, University of Pittsburgh, United States

Thank you for the opportunity to read and review this early version of Mapping Cryptic Phosphorylation Sites in the Human Proteome. The work is an interesting meta-analysis using the best quality resources for human PTMs (PhosphoSitePlus) uncovered with mass spectrometry (LCMS) and coupling these results with new versions of AlphaFold.

Overall, the study is well-conceived, and I'm encouraged to see these resources being utilized in clever new ways. I do wish to be clear regarding my own limitations as a reviewer - I am not a structural biologist, and it can be pretty tough to tell from the title whether I'm completely qualified to review a manuscript. In this case, I am not qualified to review this manuscript or methods in their entirety. Molecular dynamic and protein folding simulations, particularly the equations employed, will require the input of other reviewers with expertise in these areas.

In areas where I'm a more qualified reviewer, my main concerns revolve around the reproducibility of these results. Critical information appears to be missing, such as version information for every algorithm used in the study. What version of Seaborn? What release of Python? While the website is clever and the nomenclature edgy and fun, someone else looking to reproduce these results might not be thrilled with the fact that the source data for your manuscript is on a website under a link called “data room cryptics” which leads you to a Google Drive link. I have included recommendations for improving the reproducibility and professionalism of the manuscript in my listed numerical comments. This key point is what made me move this from a 4-star to a 3-star rating for this study.

From a biological standpoint, my only major concern with the study is my typical concern with any study using AlphaFold. How are the larger proteins being modeled? How are known limitations in error as protein sequence length increases being accounted for? While I know these tools evolve

 doi.org/10.32388/0BNAP2

rapidly, I think this should be addressed at some point in the text. For example - How is a huge protein like Titin, which can carry hundreds of phosphorylation sites, or collagen type 1A, which is at the upper limits of protein size AlphaFold can even accept and is heavily modified, accounted for in either the methods or the analysis? If Titin needed to be provided as 11 separate protein sequences for even uploading into the software for 3D modeling, that is critically important here. Aside from these, which may be oversights during the writing process, I have the following specific comments.

Minor comments:

1. It would be useful to define RSA in the text prior to the first use of the acronym. This is typically required by most journals, but this is my first Qeios review, where I suspect it is at least recommended. Actually, this may be a recurring theme in the work. Please verify that an acronym is defined before it is used in the text.
2. Version information is almost entirely absent in the methods section of the manuscript. For others to reproduce your work, this is critical to include in some way, as these tools inevitably evolve between versions. Please update the text to include this information. Seaborn, Spectrus, etc., also provide what version of Python was used and where it was executed. If
3. Supplemental Figure 2A - I'm not sure a dot plot is the most effective plot to convey the data the authors intend. A secondary plot may make this more clear to a reader or reviewer.
4. Supplemental Figure 4 - This may also be a bad match of data to plot. The density of data clustering at/near 0 is hard for me to match to the correct data point. Values observed in the non-zero space are compressed here and difficult to both observe and interpret.
5. I like the darkproteome website, as long as I don't look too long at the AI art. Unfortunately, I can't approve of the use of a Google Drive for source materials for a publication. These can be easily altered or lost. I recommend that these files be placed somewhere more typical. GitHub or FigShare repositories can be published with version control (GitHub permanently locked through services such as Zenodo). These will reduce confusion when others are attempting to reproduce your work and - for example - read this version of the manuscript and skim the later finished/published versions. The generation of DOIs that can be included within the manuscript is also important.
6. While interesting, “data room cryptics” is not the clearest description for how to access the source data for a manuscript. I strongly recommend this is changed in a way that makes it easier - rather than more confusing - for someone to reproduce the steps of your work.

7. Supplemental Table 5 appears to be missing. It may be possible to add those data to Supplemental Table 4?

I sincerely think this is an interesting modeling study, and there is a place for these kinds of bold meta-analyses. The point of writing up a scientific publication is for you to describe your findings or the merits of your hypothesis testing for other scientists to test and challenge. To get this study to the next step, you'll need to add your software versions, clearly document your settings, and provide them in a form that is easy to access, review, discuss, and reproduce. And explain how you got Titin (and the other annoyingly huge and heavily modified proteins, AHNAK is another good one) into the software and whether they match your models.

Declarations

Potential competing interests: No potential competing interests to declare.