Dear Professor Rita Casadio, we thank you and the reviewers for the time spent to review this manuscript. The remarks provided have been both useful and enlightening for us in our pursuit to perfect our manuscript. All responses are marked in red to explain what revisions have been made.

## Reviewer #1:

I found the paper **a bit difficult to read,** possibly because of the way it's **structured.**

We tried to address this by restructuring the manuscript, putting the Materials and Methods section right after the introduction. Hopefully, this will make the acronyms (see below) to be explained to the reader before they are used.

Plenty of **acronyms** crop up very rapidly in the Results section and one has to bounce back and forth to the Methods section to try and mine them. E.g. **IDDT** (with which I wasn't familiar) appears right away, keeps coming up, and then finally somewhere in the methods there is a citation next to its **24th mention** (the citation being the only explanation). I wonder whether having the citation the first time "IDDT" appears after the abstract wouldn't have helped, possibly alongside a 5-word description of what it is.

AA20 ED would have also benefited from being briefly described with words.

These issues are addressed by providing citations upon the first occurrence of each acronym as well as short descriptions of what the main acronyms (lDDT and AA20 ED) entail the first time they appear in the Results section. Partly these issues were also resolved by restructuring the manuscript, as described above.

I'm also not that sure why there is a description of AA2, AA3, etc. when AA20 is the only one that is referenced in the results.

Thanks for pointing this out. This issue was addressed by referring more clearly to all the scores (both sequence: AA2, AA3, AA6 or AA20 and structural: lDDT, RMSD, DIFFSS, DIFFC, DIFF\_ACC or TMscore) in the results section (see Comparing structural scores and sequence cardinalities). The reason to use analyzed in the supplementary material.

## Reviewer #2:

The study is interesting and suited to the journal. **The method should be more extensively described and validated**. The paper **lacks clarity in some section**. More specifically:

- It is not clear how the dataset was built: "From each family (H-group), 2-15 random entries were selected, taking the maximum possible number.

**"The maximum under which respect?**

We do understand that this was a bit unclear. We have reformulated it to up to 15 entries depending on the size of the H-group.

Why for the "Fold dataset" just **one random pair** has been selected? And **is random selection the preferable** with respect to a preliminary analysis, for example, of sequence similarity?

**This is addressed in the “fold set” in the Materials and Methods section:** Only one random pair was selected per H-group to obtain distant representatives, as compared to selecting more pairs and thus more similar homologs. Random selection was performed to ensure no selection bias is introduced for the H-group representatives. The limit of 15 pairs per H-group is also used to minimize bias from large families.

- How is conformational diversity of the same protein in different PDB entries (due for example to apo- and holo-forms or to presence of ligands) considered? To which extent is it fully taken into account in the variability estimates?

**The issue of conformational diversity due to e.g. apo- holo-forms is addressed by the following (see also Investigating the impact of ligand binding, methods section):**

To investigate the impact of ligand binding on the sequence-structure relationship we first select all ligands for each PDB ID in the Broad dataset according to PDMSum [(Laskowski et al. 2018)](https://paperpile.com/c/0nvX3q/Y7Sm). We then create one set where the PDB IDs in each domain pair do not share ligands, and one where they share at least one ligand. The sequence-structure relationships of these two sets are then compared in terms of running averages and point distributions relative to the complete Broad dataset and each other.

**A figure and Pearson correlation coefficient are produced (see Methods section and Figure S5):**

The pairs with at least one shared ligand and those without shared ligands from the Broad dataset are compared in terms of AA20 ED and lDDT score in Figure S5. As can be seen, the point distributions between these two sets are almost identical and the Pearson correlation coefficient between the running averages is 0.99.

**This is also discussed (see Discussion):**

Conformational diversity due to e.g. apo- holo-forms (enzymes without and with essential cofactors respectively ) may affect the sequence-structure relationship observed. The sequence-structure relationships between domain pairs with shared and different ligands are therefore compared with nearly identical results (the Pearson correlation coefficient is 0.99 between the running averages) regardless of ligand sharing or not (see Figure S5).

Certainly, other factors could also cause differences, but we have not been able to detect any that cause the average curve to differ substantially.

- The predictor is very shortly described. In particular, I'm not sure the provided details are sufficient to reproduce it. Moreover, it is not clear whether the hyperparameter have been set on test sets. An independent evaluation set (or a way to prevent information leaks) should be added.

Thanks for pointing this out. We have tried to describe it in more details and the sharing of all code used for this study ensures its reproducibility. Regarding the tuning of the hyperparameters, it is clarified that these have been set using only data from the training sets in the five-fold cross-validation procedure. Evaluation sets are already present, which was further explained by elucidating further the meaning of the five-fold cross-validation procedure used (see Machine Learning, Material and Methods section).

- Different measures of sequence and structural similarity are available. Authors use IDDT and ED and only a partial clarification of the reasons of the choice are given in the discussion session. For sake of clarity, I suggest to anticipate, at least in part, the issue and to further explain why these are the most suited measures. In the discussion, it would be useful to argument to which extent the results could be robust to other measures.

We address this issue by adding in the introduction:

In addition, many other measures for assessing structural similarity based on local interactions, including lDDT [[18]](https://paperpile.com/c/0nvX3q/PloOo) and CAD-score [[19]](https://paperpile.com/c/0nvX3q/pJ0j), have been developed recently. These methods are often more useful than RMSD as they are less sensitive to structural rearrangements.

And in the discussion:

To ensure the results obtained are robust, we evaluate the sequence and structure similarities with other structural scores (lDDT, RMSD, DIFFSS, DIFFC, DIFF\_ACC or TMscore, see methods section) and sequence cardinalities (AA2, AA3, AA6 or AA20, see methods section) for both initial sequence and structural alignments. As can be seen in Figure S2, all score combinations result in very similar sequence-structure relationships, yielding support for the findings here. lDDT and AA20 ED could thus easily be replaced by other scores in a retrained predictor, as can be seen in the similarity of lDDT, RMSD, DIFFC and DIFF\_ACC against AA20ED.