Author Correction: Sensitive inference of alignment-safe intervals from biodiverse protein sequence clusters using EMERALD

Following publication of the original article [1], an error was identified in the left plots of Figure 3a and b, in the Results section, under the heading *Benchmarking and analysis of stable versus unstable protein structural features that are encoded by alignment-safe windows.*

The corrected figure is the following (with changes in the caption in **bold**):

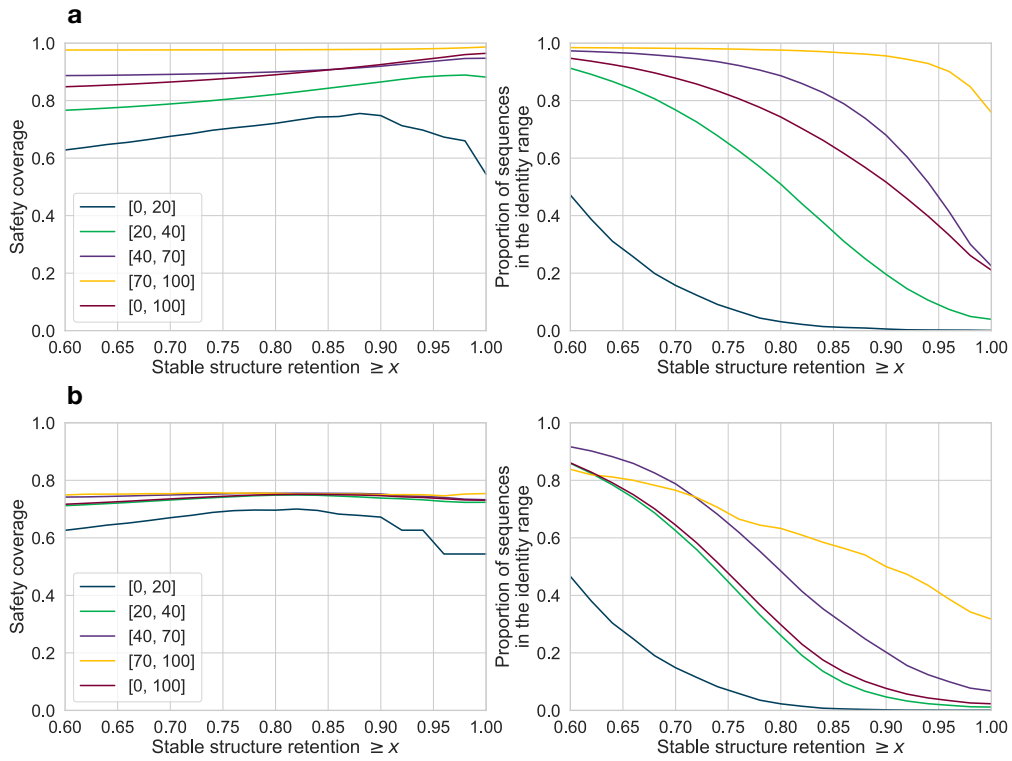

Figure 3: For a given threshold x, safety coverage of the sequences whose stable structure retention is at least x (on the left), and the proportion of the sequences whose stable structure retention is at least x (on the right). The results are split into different identity ranges. **a** All 396k sequences are included. **b** Restricted to sequences whose safety coverage is at most 80%. The dark red curves, which cover sequences of all identities, show in **a** that around 20% of the sequences (right plot) have a stable structure retention of 100% and only **96%** of safety coverage (left plot), while the purple curves, which cover sequences from the identity range 40% - 70%, show in **b** that 10% out of the sequences with at most 80% safety coverage inside this identity range (right plot), have a stable structure retention of 100% and only a safety coverage of around **73%** (left plot).

This error arose from an error in the script creating the plots, and the data that the figure was based on

remains correct. We would like to change the following paragraph under the same heading, discussing these plots:

> "Strikingly, in Fig. 3a we can see that for about 20% of all sequences, or of all sequences in the identity range $[40\%, 70\%]$ (dark red and purple curves, respectively on the right for $x = 1.00$) EMERALD retains *all* their stable positions, with a safety coverage of only 15%. This illustrates that, in contrast to optimal alignment approaches, for lower identity bounds (dissimilar sequences) EMERALD manages to reveal structurally conserved intervals. Similarly, EMERALD achieves a stable structure retention of 80% for around 75% of all the sequences with a safety coverage of around 65%. In other words, EMERALD declares 65% of the sequence as safe, capturing 80% of all stable amino acids. This result further shows that by relaxing the stable structure retention criteria from 100% down to 80%, EMERALD is able to reduce the sequences to their safety intervals from full length to 65%. In Fig. 3b, we analogously restrict the sequences only to those of safety coverage of at most 80%. Here, for example, the purple curve shows that around 10% of the sequences in the identity range 40% - 70% can be reduced to a safety coverage of nearly 5%, without losing the stable structural retention constraint of 100%. This suggests that EMERALD embraces the sequence diversity to narrow down the structurally-conserved context of a sequence."

This paragraph should read (with the changes in **bold**):

> In Fig. 3a we can see that for about 20% of all sequences, or of all sequences in the identity range $[40\%, 70\%]$ (dark red and purple curves, respectively on the right for $x = 1.00$) EMERALD retains *all* their stable positions, with a safety coverage of **95%-96%**. Similarly, **in the identity ranges** $[20\%, 40\%]$ **and** $[0\%, 20\%]$ **(green and dark blue curves),** EMERALD achieves a stable structure retention of **at least 95% and 75%** for around **10%** of the sequences, **respectively, each** with a safety coverage of only **86% and 67%**. This illustrates that, in contrast to optimal alignment approaches, for lower identity bounds (dissimilar sequences) EMERALD manages to reveal structurally conserved intervals. In Fig. 3b, we analogously restrict the sequences only to those of safety coverage at most 80%. Here, for example, the purple curve shows that around 10% of the sequences in the identity range $[40\%, 70\%]$ can be reduced to a safety coverage of **73%**, without losing the stable structural retention constraint of 100%. This suggests that EMERALD embraces the sequence diversity to narrow down the structurally-conserved context of a sequence.

The conclusions in this discussion do not change, and also the primary results and conclusions in the paper still hold.

References

[1] Andreas Grigorjew et al. "Sensitive inference of alignment-safe intervals from biodiverse protein sequence clusters using EMERALD". In: *Genome Biology* 24.1 (2023), p. 168.