THE UNIVERSITY of EDINBURGH
School of Chemistry

Joseph Black Building
David Brewster Road
Edinburgh EH9 3FJ
United Kingdom

Tel +44 (0) 131 651 3309
Fax +44 (0) 131 650 6453

antonia.mey@ed.ac.uk
www.chem.ed.ac.uk

Edinburgh, 26/05/2022

Dear Michael Thoennessen and the rest of the PRE editorial team,

Please find attached the manuscript "What geometrically constrained protein models can tell us about real-world protein contact maps" by *Nora Molkenthin, J. Jasmin Güven, Steffen Mühle and Antonia S. J. S. Mey,* that I wish to be considered for publication as a research article in Physical Review E. You can find a version of this also on the ArXiv preprint server: https://arxiv.org/abs/2205.09074. All data to reproduce the work including analysis scripts can be found in this GitHub repository.

In this paper, we investigate how the distribution of amino acid distances $\mathrm{P}(s)$, the distance between connected amino acids in a folded protein, behaves in both real world proteins and geometrically constrained protein models (Molkenthin et al PRL 117, 168301 (2016), Molkenthin et al. PloS One 15, e0229230 (2020)). Key findings are:

- We derive an analytical approximation for the amino acid distance distribution from the geometrically constrained protein model introduced in Molkenthin et al PRL 117, 168301 (2016). This analytical approximation improves on a heuristic of $\mathrm{P}(s) \approx s^{-1}$ introduced by Bartoli *et al.* [23].
- We show that the approximation fits simulated data of the 2D and 3D versions of the geometrically constrained protein models.
- We use the analytical approximation to fit protein structures taken from AlphaFold2 (deep learning model for protein structure prediction) and the protein databank (PDB – containing experimental structural protein data) for different protein sequence lengths investigating the scaling with sequence length and how well the approximation fits to these real-world data sets. We used over 400,000 protein structures in our analysis. With this, we provide an analytical way of modelling protein amino acid distances for different protein sequence lengths and show that protein sequence, on average, does not influence the distribution of the amino acid distance in biologically active proteins.

I believe our work fits well into PRE, addressing multiple remit areas i.e. Biological Physics, Polymers and Computational Physics. In particular, the analytical approximation may be of interest to readers working in the area of protein and polymer modelling and may be used as a way of modelling realistic protein chains. In addition, we provide a very thorough analysis of secondary structure data of AlphaFold 2 and PDB data providing valuable insights into the realms of applicability of AlphaFold2 structures in other research contexts.

The authors declare no conflict of interest and all funding has been acknowledged in the appropriate sections. J. Jasmin Güven and Nora Molkenthin contributed equally to the work. I am looking forward to hearing from you soon with some feedback on the manuscript.

Kind regards,

Dr. Antonia Mey
Chancellor's Fellow