Lydia E. Kavraki
Noah Harding Professor of Computer Science and Bioengineering
Professor of Electrical and Computer Engineering
Professor of Mechanical Engineering
Director, Ken Kennedy Institute

March 11, 2023

To: Briefings in Bioinformatics

Dear Editor,

Please find as an attached file the manuscript titled "***EnGens: a computational framework for generation and analysis of representative protein conformational ensembles***" (Authors: Anja Conev, Maurício M. Rigo, Didier Devaurs, Andre F. Fonseca, Hussain Kalavadwala, Martiela V. de Freitas, Cecilia Clementi, Geancarlo Zanatta, Dinler A. Antunes and Lydia E. Kavraki) that we are submitting for publication in *Briefings in Bioinformatics*.

Recent progress in machine-learning methods such as AlphaFold and ESMFold for protein structure prediction (PSP) has led to an increase of available protein structure data. These developments are driving the field of protein structure analysis closer to an era of big data. However, single protein conformations predicted by PSP models do not paint the whole picture. Proteins exist in complex environments and exhibit flexible behaviours to achieve various functions. In this novel work, we provide a comprehensive computational protocol for generating and analyzing representative conformational ensembles – EnGens. First, we provide an overview of existing unsupervised learning methodologies that can help tackle this problem. Second, we collect and combine these methodologies under the EnGens automated pipeline. EnGens is readily available for entry-level users through pre-packaged Docker image, Python package and interactive Jupyter Notebooks. Finally, we showcase the power of EnGens on relevant examples from the literature where we give better and faster insights into protein flexibility than previous manual curation. EnGens ensembles can be useful for many downstream tasks related to drug discovery such as molecular docking and drug-target interaction prediction. Additionally, EnGens can serve as a platform for further algorithmic development. Overall, we see the EnGens pipeline becoming part of many new efforts to utilize the structural data generated by novel PSP tools.

Some argue that PSP has reached the single-structure frontier and that conformational ensembles are the future of structural biology (Lane, 2023). The most recent PSP competition (CASP15) introduced a new category for modeling conformational ensembles. With this, we believe that the computational protocol described in this work is timely and of interest to Briefings in Bioinformatics. We have not had any prior discussions with a BIB Editorial Board Member about the work described in our manuscript. We confirm that this work is original and has not been

Rice University • Department of Computer Science - MS 132 • P.O. BOX 1892 • Houston, TX 77251
6100 Main St. • Houston, TX 77005 • Phone: 713-348-5737 • Email: kavraki@rice.edu • http://www.kavrakilab.org

published elsewhere, nor is it currently under consideration for publication elsewhere. We have no conflicts of interest to disclose.

We hope you find our work of interest and look forward to hearing from you,

Sincerely,

Lydia E. Kavraki