Journal of Population Therapeutics & Clinical Pharmacology

RESEARCH ARTICLE
DOI: 10.53555/f15twe94

A BIOINFORMATICS MODULE FOR UNDERGRADUATE BIOCHEMISTRY: LEVERAGING ALPHAFOLD2 TO TEACH PROTEIN STRUCTURE PREDICTION

Dr C Rathiga*

*Associate Professor, Deportment of Biochemistry, Konaseema Institute of Medical Sciences & Research Foundation, Amalapuram, Andhra Pradesh, India

ABSTRACT

AlphaFold2 uses artificial intelligence to predict the detailed three-dimensional shape of proteins using information from their amino acids. Because it is very accurate and efficient, plus it benefits many biochemistry research areas, protein structure modeling should be featured in undergraduate biochemistry courses. An instructional example was produced to present AlphaFold2 to students in an advanced biochemistry laboratory. This module focused on helping students build the models of proteins produced by the genome of the recent global outbreak, for which there were no published, experimental structures. We sought to assess how the module changed student knowledge in biochemistry and to make it possible for educators at all bioinformatics levels to add AlphaFold2-based exercises into their classes.

Key words: AlphaFold2, Protein Structure Prediction, Artificial Intelligence, Undergraduate Biochemistry Education, Bioinformatics Integration

INTRODUCTION

Structural biology is a key topic included in most biochemistry books used in lecture courses. Protein structure helps researchers with drug development, engineering proteins, designing medicines and using protein structure in biotechnology. Most often, scientists use X-ray crystallography, nuclear magnetic resonance or, in recent years, cryo-electron microscopy to investigate protein structures [1-3]. Previous studies report that these techniques are commonly applied, but instruction on them is mostly offered for graduate students. The complexity and length demanded for these results make them unfit for a single short module or class. Advances in computing changed the way DNA sequences were assembled from data acquired by running them through gels and there is now a push to use computers to address the problem of figuring out protein shapes. Since 1994, the protein-prediction methods are regularly reviewed during the Critical Assessment of Structure Prediction (CASP) competition [5]. In 2020, during CASP14, the field made its major first step as the AI company DeepMind's AlphaFold2 predicted the structure of proteins with >90% accuracy [5,6]. Despite previous barriers, AlphaFold2 successfully predicts almost perfect structures for proteins with sequences similar to those known, as well as for those whose sequences don't resemble any that have been observed [6]. This detailed the key part played by machine learning in a worldwide, free-design approach [7] which did not require a prior reference as the plan for developing models. AlphaFold2 has clearly had a big and ongoing impact on biochemical research. In the last few years, it has helped to study systems that exceeded the abilities of traditional experiments, including the cytoplasmic ring within the nuclear pore complex [8]. What's more, virtual screening in drug discovery at pharmaceutical companies can use the structural models generated by AlphaFold2. Nonetheless, we should review the obtained models carefully before reaching firm conclusions [9]. We developed a teaching module to train students in a senior-level biochemistry laboratory course on the Beta Fold protein structure software AlphaFold2. AlphaFold2 makes a great bioinformatics teaching tool since anyone can use it just by clicking. It executes using a script that handles all tasks for generating a 3D protein model, so only minor customization and little programming knowledge are needed. Interpreting the results from AlphaFold2 helps students apply important ideas about protein folding and structure they have already learned. All these features show that AlphaFold2 can be successfully included in a biochemistry curriculum at the undergraduate level. The AlphaFold2 module was delivered to 64 undergraduate students during two 3-hour class sessions. Standardized CLASS Field questions from the Colorado Learning Attitudes about Science Survey (CLASS) were used for both students and instructors to assess the module. Educational studies have used the CLASS method for analyzing student views about physics, chemistry, biology and computer science courses [10–13]. As an example of AlphaFold2's uniqueness, the scientists used recent sequences from COVID-19 virus (isolate ON563414.3) to generate 3D protein models. Students needed to look at the predicted structures with qualitative and quantitative measures. Lastly, they used a worksheet (S1 File) to review their structures and assign functions to them depending on domain homology found in the NCBI Conserved Domain Database [14].

METHODS

Ethics Statement

Approval for student and expert surveys was given by the Institutional Review Board under protocol IRB2023-0423M. Every effort was made to keep the survey anonymous for all participants. No name or contact information was obtained from the respondents and only non-involved teaching assistants gave the surveys. Any participant was able to leave the study by true-or-false during the data collection period and their responses would be excluded from the dataset.

Building of the AlphaFold2 Modul

Senior students in biochemistry and genetics, along with some juniors, take the Biochemical Techniques I course as a 15-week spring semester class. This course covers two sections of 64 students who meet together twice per week for sessions lasting 2 hours and 50 minutes. The experiments in the course focus on how charge changes the activity and structure of ribonuclease Sa [21–24]. Upon designing their experiment, students change part of the protein's sequence to alter the charge at physiological pH, induce the mutant's expression, purify it via chromatography, verify purity on a gel and confirm the mutated protein with tandem mass spectrometry. Subsequently, performance and stability are checked and molecular modeling is completed using ChimeraX. All through the semester, students save their discoveries in a style similar to scientific articles. Normally, during the final 3 to 4 weeks of the course, expert lecturers are brought in to discuss present topics in biochemical research. We all had two class periods in the last four weeks to try out AlphaFold2 ourselves. There were different goals for these two sessions (Fig 1A). In the first session, students were introduced to AlphaFold2, including the basics of protein structure, its role in life science, methods of studying protein structure, the early years of protein structure predictions through CASP events every two years and how to use bioinformatics scripts and file formats in computing all of this on the university's HPRC cluster [25]. At the second meeting, one week apart, we looked at the proper ways to assess model confidence using common measures from the field of protein structure prediction. Tags showed students how to download their results and look at 3D structures with the open-source software ChimeraX [26]. Before this module, students finished a molecular modeling activity where they downloaded ribonuclease Sa from the PDB and looked at the structure using ChimeraX. We changed the previous activity, adding a new worksheet (S1 File) designed for beginners which explores basic ChimeraX controls and also teaches how to analyze bond lengths and view a protein's surface. They were given the additional assignment to use CD-BLAST [14] and a selection of alignment tools to guess the tasks of their assigned proteins.

Estimation of the shapes of viral proteins

During this research, the viral genome contained 190 genes and matching amino acid sequences were obtained from GenBank at NCBI (ON563414.3). Unique protein sequences were allotted to the students during the simulation. To create the 3D structures, students applied AlphaFold2 (v2.1.1) to their asigned proteins. Since we found default parameters to be inappropriate, all input for AlphaFold2 was provided manually when running S3. I ran a complete database search (—db_preset = full_db) and chose the monomer_ptm model (—model = monomer_ptm) to measure the errors of the calculated alignments during the prediction. All the prediction jobs were hosted and ran using the university's HPRC cluster. AlphaFold2 generally runs its last steps faster with NVIDIA GPUs, but since only four nodes met its demands, every job was assigned to CPU nodes and took much longer. Each job was given 24 CPU cores, 160 GB RAM and a limit of 24 hours before it was cut off by the job scheduler. Most predictions are finished in around 4 hours. Students downloaded the model that received the highest rank for analysis during module session two.

Studying how far the analysis goes and the number of times each genome is sequenced

Every student evaluated their model confidence by observing three key metrics produced by AlphaFold2. Graphs for these metrics were created by using a custom script. With the help of the AlphaFold2 wrapper script, a search of many metagenome databases is conducted to create a multiple sequence alignment of the target which makes the predicted structure more accurate than the experimental results [6,27]. We counted and measured the percentage of similar homologous sequences found in the "features.pkl" file from the AlphaFold2 output. The scripts were modified to function with Google Colab Jupyter notebooks, so students don't have to configure their own software environments (S4 File). Heatmap graphs are created by this script to represent the relation between residue indices on the x-axis and homologous sequence counts on the y-axis (Fig 3B, 3F and 3J). Homologous identity for each amino acid is color-coded from no similarity at the bottom of one to identical at the top. The values come from the intensive search process used during MSA generation [28]. The number of matching sequences per amino acid is also pulled out with a black line and more homologous sequences suggest better and more confident predictions. Research beforehand suggests that if a model has 100 or more homologous sequences at a residue, it is considered a "good" model [29]. Our assessment of the data (Fig 3) indicates that often, where there are less than 30 homologues, we can notice structural ambiguities and gene-based inconsistencies.

pLDDT values are used to analyze the confidence of the local model.

Many scientists prefer the LDDT, since it compares models without having to use an original structure and gives a measure of the errors between the models and the predicted one [30]. The stepwise method for calculating LDDT is hard, but its role as a measure of suitable local structure was explained to students in the initial module lesson. Confidence in the local 3D placement of atoms is indicated by pLDDT—the greater the pLDDT, the more accurate the model is like the actual structure found in experiments. The value is stored in result_model_#ptm_pred#.pkl for every model in each AlphaFold2 run and you can plot it using alphaPICKLE [31]. The average pLDDT for every residue is displayed in the scatterplots (Fig 3C, 3G, 3K) and is further highlighted by color to help see the results more clearly. According to EMBL-EBI's AlphaFold2 database, regions with a pLDDT >90 are accurate, those between 70 and 90 are accurate for the main chain, regions with a pLDDT between 50 and 70 are not precise for side chains and <50 means it is a disordered area with nothing to interpret [6,32]. Regions with structural disorder could be areas where biomolecules change shape and cannot be identified using most laboratory techniques or they could simply be areas where the body can flex or shift physiologically.

Looking at the Expected Errors for the Both Individual and Stack Relative Models

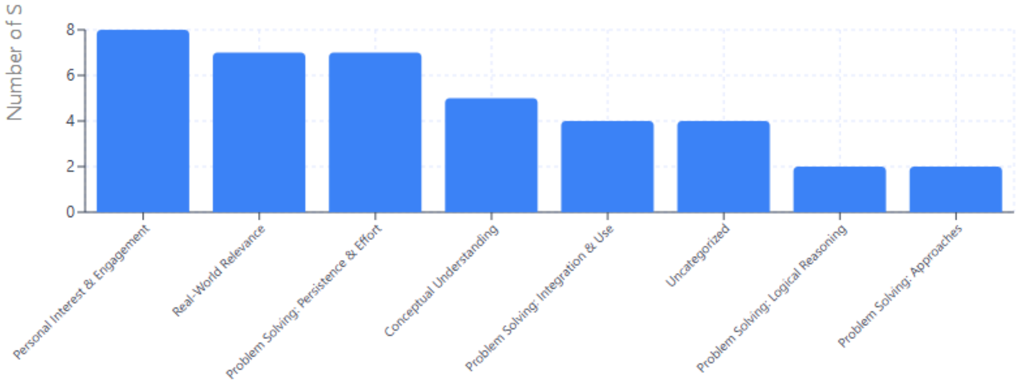
In addition, the placement of different amino acid residues compared to the reference data in the entire model must be carefully examined. By bringing together the model and its copy, AlphaFold2 sees how well the residues correspond based on Cα, N and C atoms to measure pAE [32]. Low pAE measurements show that the ordering of residues in space is well supported. Multi-domain proteins or protein complexes, when modeled by AlphaFold2, benefit especially from the improved accuracy of LDDT as a metric. AlphaPICKLE can pull out information and plot key values found in the output folder of AlphaFold2 predictions. Be sure to set the parameter in the model to “monomer_ptm” in order to calculate pAE.

RESULTS

Table 1. Survey statements organized by CLASS category.

Statement Numbers	CLASS Category
1, 3, 7, 9, 11, 13, 15, 21	Personal Interest & Engagement
3, 9, 11, 14, 17, 23, 29	Real-World Relevance
4, 6, 20, 30	Problem Solving: Integration & Use
8, 19	Problem Solving: Approaches
10, 14, 19, 22, 25, 27, 30	Problem Solving: Persistence & Effort
16, 25	Problem Solving: Logical Reasoning
12, 18, 23, 26, 28, 31	Conceptual Understanding
5, 22, 24, 26	Uncategorized

Figure.1 CLASS Category Distribution of Survey StatementsColorado Learning Attitudes about Science Survey (CLASS) statement categorization



To conduct this study, statements from the CLASS-BIO instrument were chosen and adjusted so that the survey was applicable to biochemistry. The data was divided into several groups to review engagement and how students handle biochemistry. The ratings in Personal Interest & Engagement show how much students are interested in biochemistry. In those concepts, we check if students believe the information is connected to and useful in real-world and current problems. Through “Problem Solving: Integration & Use,” students work on using different topics to solve recently introduced biochemical problems. This section is followed by “Approaches to Problem Solving,” discussing strategies used by students to deal with task they have never encountered before. In order to solve problems well, students have to be determined and work hard and the “Problem Solving: Persistence & Effort” section shows their progress. In “Problem Solving: Logical Reasoning,” the skills necessary to use logical reasoning when doing science are examined. Under “Conceptual Understanding,” students are evaluated on their connections among various concepts and how much knowledge from biochem they can retrieve from memory. A small number of statements are not tied to a specific meaning, allowing students to interpret them differently. By using this method, teachers can discover how each student understands and learns about biochemistry.

DISCUSSION

Because biochemical research now depends on bioinformatics, educators should include up-to-date educational modules for undergraduate students. We developed AlphaFold2 as a module that can be easily integrated into a biochemistry lab or classroom, even if the instructor lacks thorough bioinformatics skills. AlphaFold2 was picked because the method is open and accessible and its results can be seen in 3D and interpreted using typical biochemical knowledge of protein function. We building a module consisting of two sections on protein structure prediction, showing students how to browse the Protein Data Bank, predict the structure of any unknown protein and analyze its quality using sequence homology, pLDDT and predicted aligned error (pAE). According to the responses, about 64 students seemed interested and enthusiastic about discovering this new method in biochemistry. A comparison of surveys gathered before and after the module, as well as expert analysis with modified CLASS questions, suggested that learning about a pressing global health topic improved how students understood and applied their classroom lessons to practical use. Students examined protein structural prediction and how to use AlphaFold2 to look at possible 3D structures using three confidence measures. Most senior students did not say they were eager to take advanced bioinformatics courses right away, but they recognized its value in biochemical research and saw how it could be helpful to them in the future. Even though engagement was great, the effect on students' decisions about which courses to take could be modest, given that many of them are close to finishing their studies. In future, bioinformatics course surveys should look at how long it takes students to finish their studies. Based on this pilot study, we believe changes could improve both students' learning and their enjoyment. Letting each student predict a larger number of protein structures might improve their ability to remember which is generally supported by repetition [19]. With the structure-related data ranging from standard to more abstract, extra time spent analyzing various data and playing with models should help students feel more comfortable. If the module ran for a longer time, students could perform additional work such as running simulations to analyze protein binding and resemble how drug discovery is done today. We suggest InterPro [20] could be a better choice than the NCBI Conserved Domain tool [14] for functional annotation in the next iterations. This upper-level course is intended to connect and reinforce what you learned in Biochemistry Lecture by teaching you procedures you can practice yourself. The students spend around 11 weeks modifying, testing, cleansing and examining their ribonuclease Sa using methods developed by the department's professors. The last few weeks in the course cover more topics and prepare students for final reviews. Nothing needed to be changed in our course content because the material added came from supplementary resources. Yet, instructors who are aware of tight schedules might use the module to investigate proteins previously studied in the lab, by modeling their 3D structures with AlphaFold2 or accounting for ligands binding with the AlphaFold2 multimer model. For those educators missing high-performance computing clusters, we urge you to try ColabFold [18] which allows you to access AlphaFold2 on Google Cloud for free. Unlike other tools, ColabFold uses a more efficient search to generate predictions in around 30 minutes to an hour, created by our experience. Although AA3D is faster, those running the software should know it can be interrupted in some cases, especially when dealing with proteins that are too long or highly similar to other proteins. We suggest using ColabFold on a few sample proteins before students start their projects. Users of these tools can get small, simple protein structure predictions in roughly one hour and get analysis reports on MSA, pLDDT and pAE. Results can be directly downloaded to the cloud by students for future review.

CONCLUSION

The integration of AlphaFold2 into the undergraduate biochemistry curriculum represents a significant step towards aligning educational practices with the evolving landscape of biochemical research. This pilot module demonstrated that even students with limited bioinformatics experience can successfully engage with advanced protein structure prediction tools, enhancing their understanding of protein folding, structure, and function. The use of real-world, contemporary viral protein sequences provided a meaningful context that reinforced the relevance of bioinformatics in addressing current scientific challenges. Although the immediate impact on students' intentions to pursue further bioinformatics coursework was limited, likely due to their proximity to graduation, the overwhelmingly positive engagement suggests that such modules can stimulate interest and appreciation for computational approaches in biochemistry. Recommendations to expand the module, including increasing the number of proteins analyzed and incorporating additional computational analyses like molecular docking, promise to deepen student learning and retention. Additionally, free and accessible platforms such as ColabFold lower barriers for implementation across diverse educational settings, democratizing access to cutting-edge bioinformatics tools. Overall, incorporating AlphaFold2-based instruction equips students with valuable skills and perspectives that will be essential in modern biochemical research, preparing them to contribute effectively to a field increasingly reliant on computational methodologies**.**

REFERENCES

1.Maveyraud L, Mourey L. Protein X-ray crystallography and drug discovery. Molecules. 2020;25(5). Epub 20200225. doi: 10.3390/molecules25051030 ; PubMed Central PMCID: PMC7179213. [DOI] [PMC free article] [PubMed] [Google Scholar]

2.Hu Y, Cheng K, He L, Zhang X, Jiang B, Jiang L, et al. NMR-based methods for protein analysis. Anal Chem. 2021;93(4):1866–79. doi: 10.1021/acs.analchem.0c03830 [DOI] [PubMed] [Google Scholar]

3.Peplow M. Cryo-electron microscopy reaches resolution milestone. ACS Cent Sci. 2020;6(8):1274–7. doi: 10.1021/acscentsci.0c01048 [DOI] [PMC free article] [PubMed] [Google Scholar]

4.McLaughlin KJ. Developing a macromolecular crystallography driven CURE. Struct Dyn. 2021;8(2):020406. Epub 20210330. doi: 10.1063/4.0000089 ; PubMed Central PMCID: PMC8012065. [DOI] [PMC free article] [PubMed] [Google Scholar]

5.Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Applying and improving AlphaFold at CASP14. Proteins: Struct. Funct. Bioinf. 2021;89(12):1711–21. doi: 10.1002/prot.26257 [DOI] [PMC free article] [PubMed] [Google Scholar]

6.Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9. Epub 20210715. doi: 10.1038/s41586-021-03819-2 ; PubMed Central PMCID: PMC8371605. [DOI] [PMC free article] [PubMed] [Google Scholar]

7.Pereira J, Simpkin AJ, Hartmann MD, Rigden DJ, Keegan RM, Lupas AN. High-accuracy protein structure prediction in CASP14. Proteins: Struct. Funct. Bioinf. 2021;89(12):1687–99. doi: 10.1002/prot.26171 [DOI] [PubMed] [Google Scholar]

8.Fontana P, Dong Y, Pi X, Tong AB, Hecksel CW, Wang L, et al. Structure of cytoplasmic ring of nuclear pore complex by integrative cryo-EM and AlphaFold. Science. 2022;376(6598):eabm9326. Epub 20220610. doi: 10.1126/science.abm9326 ; PubMed Central PMCID: PMC10054137. [DOI] [PMC free article] [PubMed] [Google Scholar]

9.Borkakoti N, Thornton JM. AlphaFold2 protein structure prediction: Implications for drug discovery. Curr Opin. Struct Biol. 2023;78:102526. Epub 20230106. doi: 10.1016/j.sbi.2022.102526 ; PubMed Central PMCID: PMC7614146. [DOI] [PMC free article] [PubMed] [Google Scholar]

10.Adams WK, Perkins KK, Podolefsky NS, Dubson M, Finkelstein ND, Wieman CE. New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey. Physical Review Special Topics. Phys Educ Res. 2006;2(1):010101. doi: 10.1103/PhysRevSTPER.2.010101 [DOI] [Google Scholar]

11.Adams WK, Wieman CE, Perkins KK, Barbera J. Modifying and validating the Colorado Learning Attitudes about Science Survey for Use in Chemistry. J Chem Educ. 2008;85(10):1435. doi: 10.1021/ed085p1435 [DOI] [Google Scholar]

12. Semsar K, Knight JK, Birol G, Smith MK. The Colorado Learning Attitudes about Science Survey (CLASS) for Use in Biology. CBE—Life Sci Educ. 2011;10(3):268–78. doi: 10.1187/cbe.10-10-0133 [DOI] [PMC free article] [PubMed] [Google Scholar]
13. Dorn B, Tew AE. Becoming experts: measuring attitude development in introductory computer science. Proceeding of the 44th ACM technical symposium on computer science education. Denver, Colorado, USA: Association for Computing Machinery; 2013. p. 183–8.
14. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, et al. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res. 2011;39(Database issue):D225–9. Epub 20101124. doi: 10.1093/nar/gkq1189 ; PubMed Central PMCID: PMC3013737. [DOI] [PMC free article] [PubMed] [Google Scholar]
15. Madlung A. Assessing an effective undergraduate module teaching applied bioinformatics to biology students. PLoS Comput Biol. 2018;14(1):e1005872. Epub 20180111. doi: 10.1371/journal.pcbi.1005872 ; PubMed Central PMCID: PMC5764237. [DOI] [PMC free article] [PubMed] [Google Scholar]
16. Gauthier J, Vincent AT, Charette SJ, Derome N. A brief history of bioinformatics. Brief Bioinform. 2019;20(6):1981–96. doi: 10.1093/bib/bby063 . [DOI] [PubMed] [Google Scholar]
17. Wilson Sayres MA, Hauser C, Sierk M, Robic S, Rosenwald AG, Smith TM, et al. Bioinformatics core competencies for undergraduate life sciences education. PLoS ONE. 2018;13(6):e0196878. doi: 10.1371/journal.pone.0196878 [DOI] [PMC free article] [PubMed] [Google Scholar]
18. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. Nat Methods. 2022;19(6):679–82. doi: 10.1038/s41592-022-01488-1 [DOI] [PMC free article] [PubMed] [Google Scholar]
19. Zhan L, Guo D, Chen G, Yang J. Effects of Repetition Learning on Associative Recognition Over Time: Role of the Hippocampus and Prefrontal Cortex. Front Hum Neurosci. 2018;12:277. Epub 20180711. doi: 10.3389/fnhum.2018.00277 ; PubMed Central PMCID: PMC6050388. [DOI] [PMC free article] [PubMed] [Google Scholar]
20. Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar Gustavo A, et al. InterPro in 2022. Nucleic Acids Res. 2022;51(D1):D418–D27. doi: 10.1093/nar/gkac993 [DOI] [PMC free article] [PubMed] [Google Scholar]
21. Hebert EJ, Grimsley GR, Hartley RW, Horn G, Schell D, Garcia S, et al. Purification of ribonucleases Sa, Sa2, and Sa3 after expression in Escherichia coli. Protein Expr Purif. 1997;11(2):162–8. doi: 10.1006/prep.1997.0776 . [DOI] [PubMed] [Google Scholar]
22. Shaw KL, Grimsley GR, Yakovlev GI, Makarov AA, Pace CN. The effect of net charge on the solubility, activity, and stability of ribonuclease Sa. Protein Sci. 2001;10(6):1206–15. doi: 10.1110/ps.440101 ; PubMed Central PMCID: PMC2374010. [DOI] [PMC free article] [PubMed] [Google Scholar]
23. Yakovlev GI, Mitkevich VA, Shaw KL, Trevino S, Newsom S, Pace CN, et al. Contribution of active site residues to the activity and thermal stability of ribonuclease Sa. Protein Sci. 2003;12(10):2367–73. doi: 10.1110/ps.03176803 ; PubMed Central PMCID: PMC2366910. [DOI] [PMC free article] [PubMed] [Google Scholar]
24. Pace CN, Hebert EJ, Shaw KL, Schell D, Both V, Krajcikova D, et al. Conformational stability and thermodynamics of folding of ribonucleases Sa, Sa2 and Sa3. J Mol Biol. 1998;279(1):271–86. doi: 10.1006/jmbi.1998.1760 . [DOI] [PubMed] [Google Scholar]
25. Nasari A, Le H, Lawrence R, He Z, Yang X, Krell M, et al. Benchmarking the Performance of Accelerators on National Cyberinfrastructure Resources for Artificial Intelligence / Machine Learning Workloads. Practice and Experience in Advanced Research Computing. Boston, MA, USA: Association for Computing Machinery; 2022. p. Article 19. [Google Scholar]