OpenProteinSet: Training data for structural biology at scale

Gustaf Ahdritz
Harvard University
gahdritz@g.harvard.edu

Nazim Bouatta
Laboratory of Systems Pharmacology, Harvard Medical School
nazim_bouatta@hms.harvard.edu

Sachin Kadyan
Columbia University

Lukas Jarosch
Columbia University

Daniel Berenberg
Prescient Design, Genentech & Department of Computer Science, New York University

Ian Fisk
Flatiron Institute

Andrew M. Watkins
Prescient Design, Genentech

Stephen Ra
Prescient Design, Genentech

Richard Bonneau
Prescient Design, Genentech

Mohammed AlQuraishi
Department of Systems Biology, Columbia University
m.alquraishi@columbia.edu

Abstract

Multiple sequence alignments (MSAs) of proteins encode rich biological information and have been workhorses in bioinformatic methods for tasks like protein design and protein structure prediction for decades. Recent breakthroughs like AlphaFold2 that use transformers to attend directly over large quantities of raw MSAs have reaffirmed their importance. Generation of MSAs is highly computationally intensive, however, and no datasets comparable to those used to train AlphaFold2 have been made available to the research community, hindering progress in machine learning for proteins. To remedy this problem, we introduce OpenProteinSet, an open-source corpus of more than 16 million MSAs, associated structural homologs from the Protein Data Bank, and AlphaFold2 protein structure predictions. We have previously demonstrated the utility of OpenProteinSet by successfully retraining AlphaFold2 on it. We expect OpenProteinSet to be broadly useful as training and validation data for 1) diverse tasks focused on protein structure, function, and design and 2) large-scale multimodal machine learning research.

1 Introduction

Multiple sequence alignments (MSAs) comprise sets of related protein sequences with their amino acid residues in correspondence ("aligned"). MSAs encode rich information about the functional and structural features of a protein family by summarizing the (co-)evolutionary trajectory of its sequence.

MSAs are used in a wide variety of bioinformatic applications, including protein function prediction [3, 4, 5], protein language models [6, 7, 8, 9, 10], disease variant prediction [11, 12], phylogeny [13, 14], protein design [15, 16, 17], protein classification [18], and, most notably, protein structure prediction [19, 25, 26, 27, 28, 29, 30, 31, 32, 20, 21, 22, 23, 24]. Early work on the latter, culminating

37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks.

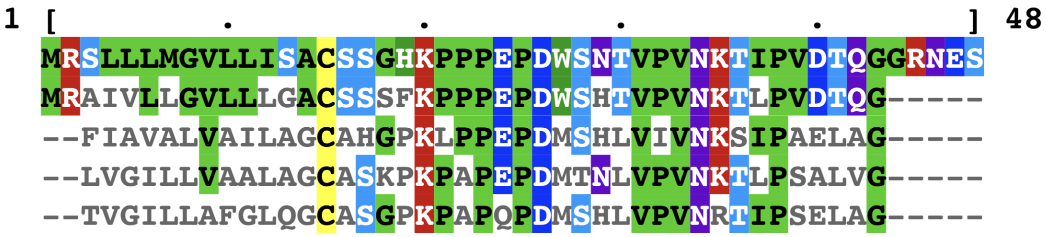
1 [. . . .] 48
MRSLLLMGVLLISACSSGHKPPPEPDWSNTVPVNKTIPVDTQGGRNES
MRAIVLLGVLLLGACSSSFKPPPEPDWSHTVPVNKTLPVDTQG-----
--FIAVALVAILAGCAHGPKLPPEPDMSHLVIVNKSIPAELAG-----
--LVGILLVAALAGCASKPKPAPEPDMTNLVPVNKTLPSALVG-----
--TVGILLAFGLQGCASGPKPAPQPDMSHLVPVNRTIPSELAG-----

Figure 1: **MSA primer**. Five rows of the OpenProteinSet MSA for PDB protein 3ZBI, chain C [1, 2]. Each row of an MSA is a protein sequence. Proteins are one-dimensional strings composed with a vocabulary of 20 amino acids—or "residues"—each represented by a letter. The target or "query" protein is given in the first row of the MSA. Subsequent rows are evolutionarily related ("homologous") proteins retrieved from a sequence database on the basis of similarity to the query sequence. To improve alignments and accommodate sequences whose length has changed over time, MSA alignment software can insert "gaps" (represented here by dashes) in or delete residues from homologous sequences. Highlights indicate conserved residues. The number of homologous sequences in an MSA ("depth") and their diversity both contribute to the MSA's usefulness.

in the original AlphaFold, achieved notable success by training models on summary statistics derived from MSAs [19, 25, 26, 27, 28, 29, 30, 31, 32, 20, 21]. More recently, large transformer-like neural networks [33] that predict protein structure by directly attending over raw MSAs came to prominence [7]. Among them, AlphaFold2 reached near-experimental accuracy for most proteins at the 14th biannual Critical Assessment of Structure Prediction (CASP) by attending over raw MSAs alongside *structural templates* of homologous proteins [23]. Follow-up work, including RoseTTAFold and the state-of-the-art protein complex structure prediction model AlphaFold-Multimer [24, 34], build on the same techniques. The dependence of these methods on sufficiently deep, diverse MSAs and close structural homologs is evidenced by the fact that they perform worst on proteins that lack them [23].

Despite the central importance of MSAs, the quantity of precomputed MSAs accessible to the research community has not kept pace with the demands of modern machine learning methods. Large models like AlphaFold2 or MSA Transformer [7] were trained on internal datasets of millions of MSAs, and the computation of various official databases of AlphaFold2 predictions [35, 36, 37] would have required hundreds of millions more. None of this data has yet been released to the public, however, and existing public MSA databases [38, 39, 40] are comparatively small and outdated. Raw sequence and structure data are available in large quantities under open licenses [23, 41, 42, 43] and there also exist several mature, open-source software suites for computing MSAs at varying levels of sensitivity [44, 45, 46]. Together, these resources are sufficient to generate MSAs at scale; indeed, they were used to create the aforementioned unreleased datasets. Nevertheless, doing so is computationally expensive. Depending on target sequence length and the size of the sequence database being searched, generating a single MSA with high sensitivity can take several hours. This effectively renders research at the forefront of protein machine learning and bioinformatics inaccessible to all but a few large research groups.

Here, we present OpenProteinSet, a large corpus of precomputed MSAs suitable for training bioinformatic models at the scale of AlphaFold2 and beyond. OpenProteinSet contains an updated reproduction of AlphaFold2's unreleased training set, including MSAs and structural template hits for all unique Protein Data Bank (PDB) chains. It also incorporates more than sixteen million MSAs, computed for each cluster in Uniclust30 [47]. From these, we identify a maximally diverse and deep subset of MSAs that are well-suited for AlphaFold2-style training runs and provide associated AlphaFold2 structure predictions.

We have demonstrated the utility of OpenProteinSet by using it to train OpenFold, a trainable, open-source reproduction of AlphaFold2 [48], achieving accuracy at parity with that of DeepMind's original model. Model parameters resulting from these experiments have been made publicly available.

Not counting these validation experiments or postprocessing, OpenProteinSet represents millions of compute-hours.

After a brief review of related work in Section 2, we provide an overview of the composition of OpenProteinSet in Section 3. Section 4 describes our retraining experiments. We conclude with a discussion in Section 6.

Sequence origin	Count (approx.)	MSA	Template hits	Structure
PDB (all unique chains)	140,000	✓	✓	Experimentally determined
Uniclust30 (filtered)	270,000	✓	✓	Predicted by AlphaFold2
Uniclust30 (unfiltered)	16 million	✓	×	×

Table 1: OpenProteinSet at a glance.

2 Related work

MSAs in structural bioinformatics: Techniques based on identifying residue-residue correlations in MSAs ("co-variation analysis") are ubiquitous in structural bioinformatics. They have existed in various forms for more than two decades [49, 50], but were initially constrained by the unavailability of sufficient protein sequence data to generate deep MSAs (*i.e.,* comprising many highly diverse sequences). With the onset of next-generation sequencing technology, exponential growth in sequenced genomes and metagenomes led to an explosion in the availability of protein sequence data.

This explosion enabled some of the first successful applications of MSA-based structure prediction methods to proteins [19, 25, 26]. To date, modern machine learning-based approaches rely almost exclusively on MSAs. The first successful models applied residual and convolutional architectures to preprocessed MSA summary statistics [27, 28, 20, 30, 31, 21, 32]. The MSA Transformer was the first to successfully apply transformers to a large corpus (26 million) of unprocessed MSAs in an unsupervised fashion [7], extending prior work on protein language models (PLMs) [10, 8, 51]. Contemporaneously, AlphaFold2 was developed to take MSAs as input to predict protein structures and is additionally trained with an unsupervised BERT-style masked MSA prediction objective [52]. The resulting model, along with its successor AlphaFold-Multimer, has been widely recognized as a revolution in protein structure prediction. Since then, protein structure prediction models that replace MSAs with embeddings from giant PLMs have emerged [22, 9, 53]. They show promise as an emerging technology, but they have so far failed to match the performance of MSA-based methods across the board, significantly underperforming AlphaFold2-based entrants on difficult targets at the most recent installment of CASP [54].

While protein structure prediction is perhaps the most celebrated use case for MSAs, they are broadly used in other areas of bioinformatics. Analogously to natural language processing, unsupervised language modeling of raw MSAs produces rich representations with broad applicability, including in protein design [16], semantic similarity inference [55], and few-shot protein function prediction, where MSA-based models outperformed comparable models trained on individual sequences alone [5]. Long before transformers, summary statistics manually derived from MSAs were already indispensable inputs for diverse tasks ranging from protein classification [18] to disease variant prediction [11, 12].

MSA software: There exists a large ecosystem of software for computing MSAs by querying large sequence databases. The commonly used programs HHMer [45] and HHblits [44] are highly sensitive, identifying evolutionarily related proteins with high recall and precision. These tools are slow and memory-intensive, however; they may run for several hours or even days to compute a single MSA. As an alternative, the efficient MMSeqs2 method trades off sensitivity for an order-of-magnitude improvement in runtime and is commonly used for fast inference with AlphaFold2, most notably in ColabFold [56]. Like the MSAs on which AlphaFold2 was trained, OpenProteinSet MSAs are computed with HHMer and HHblits for maximal sensitivity.

MSA databases: Responding to the high demand for precomputed MSAs, the community has produced a handful of public MSA repositories. ProteinNet, a repository of standardized protein data for the purposes of machine learning, includes MSAs for approximately 100,000 Protein Data Bank (PDB) protein structures released before May 2016 [40]. Earlier databases are much smaller and less diverse [38, 39]. After the initial release of OpenProteinSet in June 2022, a handful of other open MSA repositories have begun to appear, including PSP, a repository of approximately 1 million MSAs computed with MMSeqs2 [57], and a similar reproduction of about 500,000 MSAs generated according to the procedures outlined in the AlphaFold2 paper [58]. OpenProteinSet is more accurate and larger than any other MSA database.

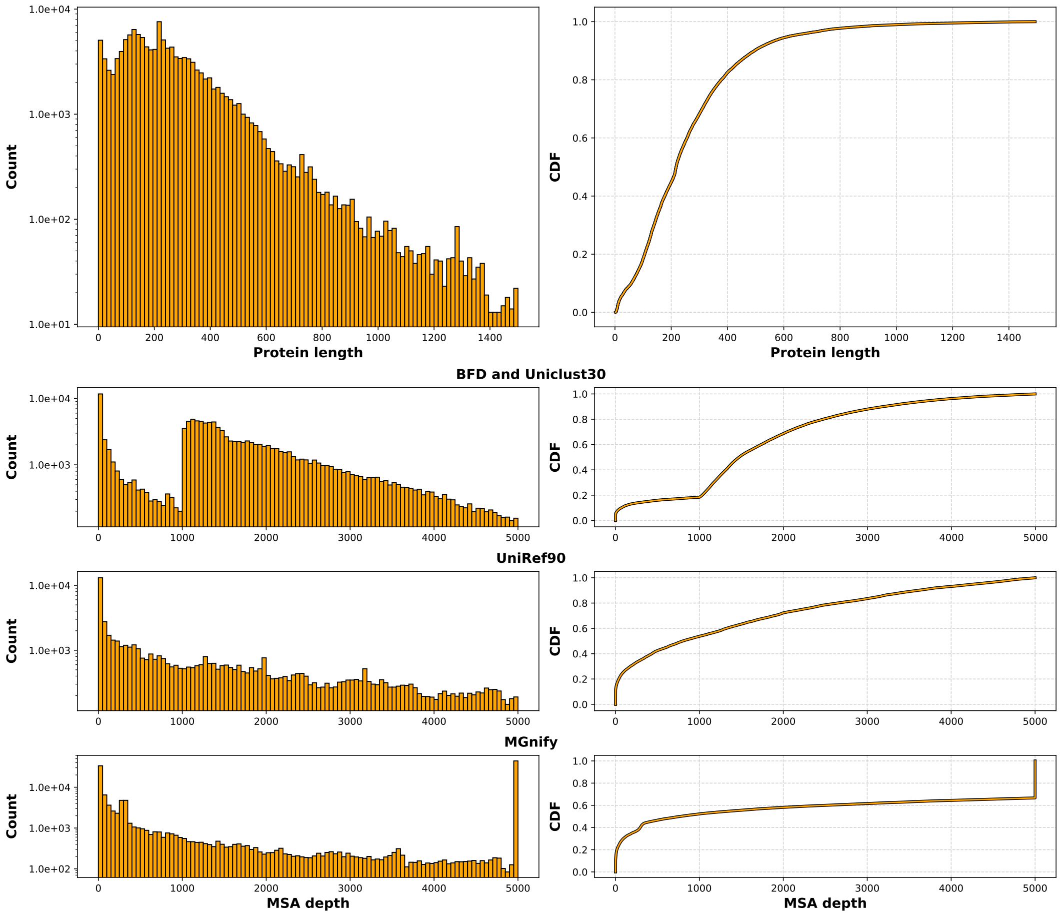

Figure 2: **PDB MSA statistics.** (First row) Number of proteins by sequence length in the PBD portion of OpenProteinSet (left) and the corresponding cumulative density function (CDF) (right). The mean length is 265; the median is 218. (Bottom rows) Depths of MSAs in the PDB portion of OpenProteinSet (left) and the corresponding cumulative density function (CDF) (right). Note that three MSAs are computed for each PDB chain in OpenProteinSet: one using BFD and Uniclust30 (top), one using UniRef90 (middle), and one using MGnify (bottom).

3 Methodology

OpenProteinSet consists of more than 16 million unique MSAs generated according to the procedures outlined in the AlphaFold2 paper [23]. This count includes MSAs for all 140,000 unique chains available in the PDB as of April 2022, immediately before the beginning of CASP15, and 16 million MSAs computed for each sequence cluster in Uniclust30 against the same database. From this latter set, we identify 270,000 maximally diverse representative clusters suitable to e.g. serve as the self-distillation set in the AlphaFold2 training procedure. Structural template hits and structure files are also available for this set and all PDB chains.

For each PDB chain, we compute three MSAs using different alignment tools and sequence databases. JackHMMer [45] was used to separately search MGnify [42] and UniRef90 [59]; HHblits-v3 was used to search the Big Fantastic Database (BFD) [23] and Uniclust30 [47]. BFD is a large sequence database prepared for AlphaFold2 that draws its approximately 2.2 billion entries from reference databases, metagenomes, and metatranscriptomes. MgNify (as of 2019) is another environmental database of approximately 300 million sequences. UniRef90 and Uniclust30 are clusterings of UniprotKB [43] proteins at 90% and 30% pairwise sequence identity, respectively, using different clustering algorithms.

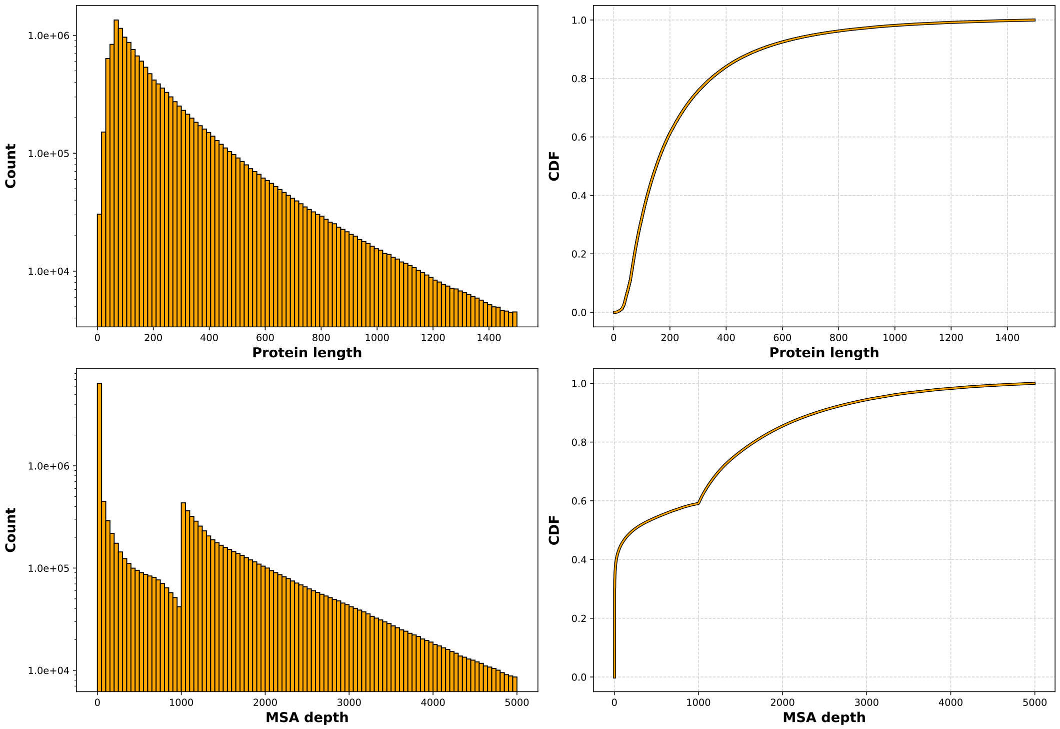

Figure 3: **Uniclust30 MSA statistics.** (Top) Number of proteins by sequence length in the Uniclust30 portion of OpenProteinSet (left) and the corresponding cumulative density function (CDF) (right). The mean length is 255; the median is 153. (Bottom) Depths of MSAs in the Uniclust30 portion of OpenProteinSet (left) and the corresponding cumulative density function (CDF) (right). The average MSA depth is 940; the median is 262.

Structural templates were identified by searching PDB70 [60] using the UniRef90 MSA using HHSearch [44]. Corresponding structures can be retrieved from publicly available PDB mmCIF files using scripts in OpenFold [48].

As in the procedure used to generate AlphaFold2's training set, we changed some of the default options of MSA generation tools. For a list of specific command-line options changed, please consult the supplementary material. One important change is that HHBlits was run for three iterations.

To generate the Uniclust30 MSAs, we performed an all-against-all search on Uniclust30 using HHblits-v3 with the same parameter settings as before. This yielded approximately 16 million MSAs, one for each cluster.

To create a filtered subset of diverse and deep MSAs, we then iteratively removed MSAs whose representative chain appeared in the greatest number of other MSAs. This was repeated until each representative chain appeared only in its own MSA. For parity with the corresponding (unreleased) AlphaFold2 set, we further removed clusters whose representative sequences were longer than 1,024 residues or shorter than 200. Finally, we removed clusters whose corresponding MSAs contained fewer than 200 sequences, leaving just 270,262 MSAs. Template hits were again computed using HHsearch against PDB70. For each representative chain in this subset, we generated structure predictions using OpenFold run with AlphaFold2 weights. Note that, unlike the hundreds of millions of AlphaFold2 predictions made available by DeepMind and EMBL-EBI [35, 36, 37], these are paired with high-quality, diverse MSAs, making it possible to use them as training data for new structure prediction models. All of the above—the 16 million unfiltered Uniclust30 MSAs and filtered-chain template hits and structure predictions—are included in OpenProteinSet.

Overall, the MSAs in OpenProteinSet represent more than four million hours of computation. Its contents are summarized in Table 1.

All MSAs are in A3M format.[1] Template hits are provided in HHSearch's HHR format, while structure predictions are in PDB format. All data is made available under the CC BY 4.0 license.

For all MSAs currently in OpenProteinSet, we used copies of UniRef90 downloaded on December 19, 2021, BFD downloaded on December 20, 2021, Uniclust30 downloaded on December 28, 2021, and MGnify downloaded on January 14, 2022. To compute templates, we used PDB70 downloaded on December 19, 2021. In all cases, we used the most recent versions of each database available at the time. As we update OpenProteinSet with new sequences, we will continually upgrade them.

We used HH-suite version 3.3.0 (commit hash `dc74ac`) and jackhmmer from HMMER3.1.

4 Experiments

To demonstrate the utility of OpenProteinSet, we used it as training data for a replication of AlphaFold2, a groundbreaking but previously unreplicated protein structure prediction network trained on raw MSAs. Our AlphaFold2 training code is implemented in OpenFold, our open-source reproduction of the AlphaFold2 training code [48].

First, we simulated the full AlphaFold2 training procedure outlined in Table 4 of the supplement to the AlphaFold2 paper. We used the PDB component of OpenProteinSet as the initial training set and our set of 270,000 filtered Uniclust30 proteins as the self-distillation set. We used a PDB cutoff of December 2021. Training was run on a cluster of 44 A100s. Given the prohibitive costs of training the full model from scratch, original AlphaFold2 weights were used as the pre-distillation model to generate Uniclust30 structure predictions.

To evaluate the resulting OpenFold weights against AlphaFold2, we computed `model_1` predictions for each currently available "all groups" CASP15 domains ($n = 90$) and evaluated them using the GDT-TS score [61]. OpenFold reached a mean score of 73.8 (95% confidence interval = 68.6 - 78.8) while AlphaFold2 reached 74.6 (95% confidence interval = 69.7 - 79.2). Confidence intervals of each mean are estimated from 10,000 bootstrap samples. OpenFold did at least as well as AlphaFold2 on exactly 50% of targets. Superimposed predictions are shown in Figure 4.

Weights from this experiment are available under a permissive license in the OpenFold GitHub repository.[2]

Next, to estimate variance from different weight initializations and other sources of randomness in training, we trained 15 models on the PDB tranche with different seeds for 10,000 initial training steps (compared to more than 75,000 in the full training run), taking advantage of the fact that OpenFold/AlphaFold2 achieves much of its final accuracy relatively quickly (as much as 90% of its final accuracy in less than 3% of the total training time). We observe very little run-to-run variability. For assessment we use lDDT-Cα [62], a commonly used accuracy measure for protein structure predictions. We found that on a validation set of 180 unseen CAMEO [63] proteins drawn over a three-month period lDDT-Cα was 0.866; the maximum value was 0.881 and the minimum was 0.848, while the median was 0.868. Our final model trained for the full duration on both the PDB and filtered Uniclust30 datasets scores 0.907 on the same validation set.

For more details on both sets of experiments, including specific hyperparameter settings, consult the OpenFold paper [48].

5 Limitations

Many centralized sequence databases are rarely updated, and while we used the most recent versions of each wherever possible, most of the MSAs currently in OpenProteinSet were computed in early 2022. Given that the number and diversity of known sequences is continually increasing, this means that OpenProteinSet—like any repository of precomputed MSAs—may age over time and need to be updated for optimal downstream performance. OpenProteinSet entries that currently have shallow MSAs or few structural homologs are particularly "vulnerable" in this regard. While we

[1] A3M is a plaintext format consisting of aligned sequences, one per line (as in Figure 1), and comment lines beginning with '>'. Gaps are represented by dashes ('-') and insertions are represented with lowercase residue letters.

[2] URL: https://github.com/aqlaboratory/openfold

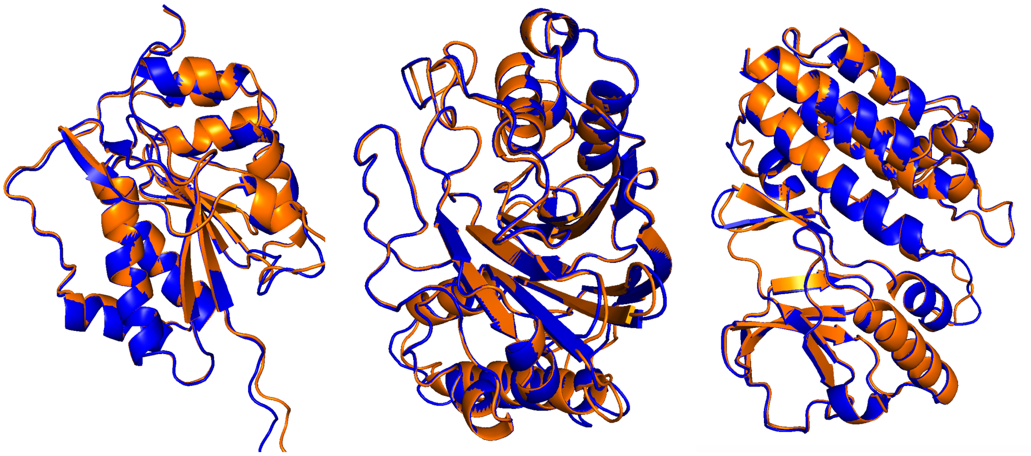

Figure 4: **OpenFold trained with OpenProteinSet reproduces AlphaFold2.** Superimposed OpenFold (orange) and AlphaFold2 (blue) predictions on three CASP15 domains: from left to right, T1109 (RMSD: 0.306), T1153 (RMSD: 0.263), and T1195 (RMSD: 0.235).

may periodically expand OpenProteinSet with new MSAs, we do not currently plan to update MSAs already in the dataset as new sequences become available.

We note too that we only evaluate OpenProteinSet on monomeric structure prediction and not other popular applications. Nevertheless, the utility of large quantities of MSAs has been firmly established in diverse settings, and we have no reason to believe that OpenProteinSet MSAs in particular will be less useful.

6 Discussion

With OpenProteinSet, we have greatly increased the quantity and quality of precomputed MSAs available to the molecular machine learning communities. The dataset has immediate applications to diverse tasks in structural biology. Below, for illustrative purposes, we highlight a handful of additional tasks and settings where we strongly expect high-quality multiple sequence alignments like those in OpenProteinSet to be immediately useful.

Protein language modeling: Unsupervised protein language models [8, 64, 22, 7, 65] have become workhorses in the bioinformatic community, as, analogously to natural language models, they encode useful biological knowledge that allows them to reason about numerous protein-related tasks. Most are trained on individual protein sequences, but MSA Transformer, a model trained on millions of (unreleased) Uniclust30 MSAs, was able to outperform conventional protein language models on downstream evaluations like protein design, and with fewer parameters [7, 16]. With OpenProteinSet, a dataset of millions of comparable Uniclust30 MSAs, it is now possible for the open-source community to experiment with similar MSA language models, perhaps even in combination with widely available single-sequence data.

Orphan proteins: One function of OpenProteinSet is to identify a large number of proteins with few or no known homologs at the time of its creation. "Orphan" proteins like these are often failure cases of models trained on protein data. In protein structure prediction, for example, MSA-based models like AlphaFold2 and RoseTTAFold are known to perform less well on proteins with shallow MSAs [23, 24]. Protein language models are slightly less sensitive to MSA depth in some cases [22], but the gap persists there as well. We expect that a large quantity of additional data on such proteins will be useful to validate and improve bioinformatic methods. Because OpenProteinSet effectively clusters sequence space, it also enables important validation experiments not possible with unclustered sequences alone, like training on one region of protein space and testing on another.

Multimodal deep learning: Beyond bioinformatics, a popular line of deep learning research studies the effects of training extremely large neural networks on data from diverse modalities. While the most commonly studied modality pairing is language and image data [66, 67, 68, 69, 70, 71], unsupervised co-training on additional modalities—including audio [72], robotics tasks [71, 73], and, indeed, raw protein sequence data—has been shown to enrich the knowledge and capabilities of models. Multimodal language models jointly trained on English text and biological sequence data have already been used to identify protein-protein interactions [74], classify adverse reactions to drugs [75], and caption molecules [76]. The multimodal scientific language model Galactica was also trained on protein sequences [77]. More indirectly, protein data often appears as a component in benchmarks for multimodal training methods. It has recently been added to DABS, a multimodal benchmark for unsupervised learning techniques [78, 79], and has been used to study multimodal scaling laws in generative models [80] and test the capabilities of pretrained language models across modalities [81]. As models become increasingly data-hungry, we believe databases like OpenProteinSet will be valuable on both of these fronts, as reservoirs of biological knowledge for generalist multimodal language models and also as tools for the empirical study of multimodal training *per se*.

Overall, we hope that OpenProteinSet will further democratize research in bioinformatics, machine learning on proteins, and beyond.

Acknowledgments and Disclosure of Funding

We would like to thank the Flatiron Institute for providing computing resources and Amazon Web Services for hosting OpenProteinSet. Individually, we would like to thank Milot Mirdita and Martin Steinegger for their valuable support and expertise.

G.A. is supported by a Simons Investigator Fellowship, NSF grant DMS-2134157, DARPA grant W911NF2010021, and DOE grant DE-SC0022199. N.B. is supported by DARPA PANACEA program grant HR0011-19-2-0022 and NCI grant U54-CA225088. M.A. is a member of the Scientific Advisory Boards of Cyrus Biotechnology, Deep Forest Sciences, Nabla Bio, Oracle Therapeutics, and FL2021-002, a Foresite Labs company.

References

[1] A. Rivera-Calzada et al. Structure of a bacterial type IV secretion core complex at subnanometre resolution. The EMBO Journal (2013), 1195–1204. DOI: 10.1038/emboj.2013.58.

[2] F. Madeira, M. Pearce, A. R. N. Tivey, P. Basutkar, J. Lee, O. Edbali, N. Madhusoodanan, A. Kolesnikov, and R. Lopez. Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Research 50.W1 (2022), W276–W279. DOI: 10.1093/nar/gkac240.

[3] S. de Oliveira and C. Deane. Co-evolution techniques are reshaping the way we do structural bioinformatics. F1000Research 6 (2017), 1224. DOI: 10.12688/f1000research.11543.1.

[4] A. J. Riesselman, J. B. Ingraham, and D. S. Marks. Deep generative models of genetic variation capture the effects of mutations. Nature Methods 15 (10 2018), 816–822. DOI: 10.1038/s41592-018-0138-4.

[5] J. Meier, R. Rao, R. Verkuil, J. Liu, T. Sercu, and A. Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. In: Advances in Neural Information Processing Systems. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan. Vol. 34. 2021, 29287–29303. https://proceedings.neurips.cc/paper_files/paper/2021/file/f51338d736f95dd42427296047067694-Paper.pdf.

[6] R. Rao, J. Meier, T. Sercu, S. Ovchinnikov, and A. Rives. Transformer protein language models are unsupervised structure learners. bioRxiv (2020). DOI: 10.1101/2020.12.15.422761.

[7] R. M. Rao, J. Liu, R. Verkuil, J. Meier, J. Canny, P. Abbeel, T. Sercu, and A. Rives. MSA Transformer. In: Proceedings of the 38th International Conference on Machine Learning. Ed. by M. Meila and T. Zhang. Vol. 139. Proceedings of Machine Learning Research. July 2021, 8844–8856. https://proceedings.mlr.press/v139/rao21a.html.

[8] A. Rives et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences 118.15 (2021), e2016239118. DOI: 10.1073/pnas.2016239118.

[9] Z. Lin et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379.6637 (2023), 1123–1130. DOI: 10.1126/science.ade2574.

[10] E. C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, and G. M. Church. Unified rational protein engineering with sequence-based deep representation learning. Nature Methods 16 (12 2019), 1315–1322. DOI: 10.1038/s41592-019-0598-1.

[11] A. J. Riesselman, J. B. Ingraham, and D. S. Marks. Deep generative models of genetic variation capture the effects of mutations. Nature Methods 15 (10 Nov. 2018), 816–822. DOI: 10.1038/s41592-018-0138-4.

[12] J. Frazer, P. Notin, M. Dias, A. Gomez, J. K. Min, K. Brock, Y. Gal, and D. S. Marks. Disease variant prediction with deep generative models of evolutionary data. Nature 599 (7883 Nov. 2021), 91–95. DOI: 10.1038/s41586-021-04043-8.

[13] K. Nguyen, X. Guo, and Y. Pan. Phylogeny in Multiple Sequence Alignments. In: Multiple Biological Sequence Alignment. 2016. Chap. 6, 103–112. DOI: https://doi.org/10.1002/9781119273769.ch6.

[14] H. Ashkenazy, I. Sela, E. Levy Karin, G. Landan, and T. Pupko. Multiple Sequence Alignment Averaging Improves Phylogeny Reconstruction. Systematic Biology 68.1 (June 2018), 117–130. DOI: 10.1093/sysbio/syy036.

[15] A. Hawkins-Hooker, F. Depardieu, S. Baur, G. Couairon, A. Chen, and D. Bikard. Generating functional protein variants with variational autoencoders. PLOS Computational Biology 17.2 (Feb. 2021), 1–23. DOI: 10.1371/journal.pcbi.1008736.

[16] D. Sgarbossa, U. Lupo, and A.-F. Bitbol. Generative power of a protein language model trained on multiple sequence alignments. eLife 12 (Feb. 2023), e79854. DOI: 10.7554/eLife.79854.

[17] V. Frappier and A. E. Keating. Data-driven computational protein design. Current Opinion in Structural Biology 69 (2021), 63–69. DOI: https://doi.org/10.1016/j.sbi.2021.03.009.

[18] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan. Protein Sectors: Evolutionary Units of Three-Dimensional Structure. Cell 138 (4 2009), P774–786. DOI: 10.1016/j.cell.2009.07.038.

[19] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. Proceedings of the National Academy of Sciences 106.1 (2009), 67–72. DOI: 10.1073/pnas.0805923106.

[20] Y. Liu, P. Palmedo, Q. Ye, B. Berger, and J. Peng. Enhancing Evolutionary Couplings with Deep Convolutional Neural Networks. Cell Systems 6 (1 2018), 65–74. DOI: 10.1016/j.cels.2017.11.014.

[21] J. Xu, M. McPartlon, and J. Li. Improved protein structure prediction by deep learning irrespective of co-evolution information. Nature Machine Intelligence 3 (7 2021), 601–609. DOI: 10.1038/s42256-021-00348-5.

[22] R. Chowdhury et al. Single-sequence protein structure prediction using a language model and deep learning. Nature Biotechnology (2022). DOI: 10.1038/s41587-022-01432-w.

[23] J. Jumper et al. Highly accurate protein structure prediction with AlphaFold. Nature 577 (7792 2021), 583–589. DOI: 10.1038/s41586-021-03819-2.

[24] M. Baek et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science 373.6557 (2021), 871–876. DOI: 10.1126/science.abj8754.

[25] D. T. Jones, D. W. A. Buchan, D. Cozzetto, and M. Pontil. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 28.2 (Nov. 2011), 184–190. DOI: 10.1093/bioinformatics/btr638.

[26] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander. Protein 3D Structure Computed from Evolutionary Sequence Variation. PLOS ONE 6.12 (2011), 1–20. DOI: 10.1371/journal.pone.0028766.

[27] D. T. Jones, T. Singh, T. Kosciolek, and S. Tetchner. MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. Bioinformatics 31 (7 2015), 999–1006. DOI: 10.1093/bioinformatics/btu791.

[28] V. Golkov, M. J. Skwark, A. Golkov, A. Dosovitskiy, T. Brox, J. Meiler, and D. Cremers. Protein contact prediction from amino acid co-evolution using convolutional networks for graph-valued images. In: Advances in Neural Information Processing Systems. Ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett. Vol. 29. 2016. https://proceedings.neurips.cc/paper_files/paper/2016/file/2cad8fa47bbef282badbb8de5374b894-Paper.pdf.

[29] S. Wang, S. Sun, Z. Li, R. Zhang, and J. Xu. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. PLOS Computational Biology 13.1 (2017), 1–34. DOI: 10.1371/journal.pcbi.1005324.

[30] J. Ingraham, A. Riesselman, C. Sander, and D. Marks. Learning Protein Structure with a Differentiable Simulator. In: International Conference on Learning Representations. 2019. https://openreview.net/forum?id=Byg3y3C9Km.

[31] M. AlQuraishi. End-to-End Differentiable Learning of Protein Structure. Cell Systems 8.4 (2019), 292–301.e3. DOI: 10.1016/j.cels.2019.03.006.

[32] A. W. Senior et al. Improved protein structure prediction using potentials from deep learning. Nature 577 (7792 2020), 706–710. DOI: 10.1038/s41586-019-1923-7.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention Is All You Need. 2017. DOI: 10.48550/ARXIV.1706.03762.

[34] R. Evans et al. Protein complex prediction with AlphaFold-Multimer. bioRxiv (2022). DOI: 10.1101/2021.10.04.463034.

[35] K. Tunyasuvunakool et al. Highly accurate protein structure prediction for the human proteome. Nature 596 (7873 2021), 590–596. DOI: 10.1038/s41586-021-03828-1.

[36] M. Varadi et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Research 50.D1 (2021), D439–D444. DOI: 10.1093/nar/gkab1061.

[37] E. Callaway. 'The entire protein universe': AI predicts shape of nearly every known protein. Nature 608 (4 2022), 15–16. DOI: 10.1038/d41586-022-02083-2.

[38] R. P. Joosten, T. A. te Beek, E. Krieger, M. L. Hekkelman, R. W. W. Hooft, R. Schneider, C. Sander, and G. Vriend. A series of PDB related databases for everyday needs. Nucleic Acids Research 39 (Database issue 2011), D411–D419. DOI: 10.1093/nar/gkq1105.

[39] S. Ovchinnikov, H. Kamisetty, and D. Baker. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. eLife 3 (2014). Ed. by B. Roux, e02030. DOI: 10.7554/eLife.02030.

[40] M. AlQuraishi. ProteinNet: a standardized data set for machine learning of protein structure. BMC Bioinformatics 20.311 (1 2019). DOI: 10.1186/s12859-019-2932-0.

[41] wwPDB Consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Research 47.D1 (2018), D520–D528. DOI: 10.1093/nar/gky949.

[42] A. L. Mitchell et al. MGnify: the microbiome analysis resource in 2020. Nucleic Acids Research 48.D1 (2020), D570–D578. DOI: 10.1093/nar/gkz1035.

[43] UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research 49.D1 (2021), D480–D489. DOI: 10.1093/nar/gkaa1100.

[44] M. Remmert, A. Biegert, A. Hauser, and J. Söding. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature Methods 9 (2 2012), 173–175. DOI: 10.1038/nmeth.1818.

[45] L. S. Johnson, S. R. Eddy, and E. Portugaly. Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics 11 (1 2010), 431. DOI: 10.1186/1471-2105-11-431.

[46] M. Steinegger and J. Söding. Clustering huge protein sequence sets in linear time. Nature Communications 9 (1 2018), 2542. DOI: 10.1038/s41467-018-04964-5.

[47] M. Mirdita, L. von den Driesch, C. Galiez, M. J. Martin, J. Söding, and M. Steinegger. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Research 45.D1 (2017), D170–D176. DOI: 10.1093/nar/gkw1081.

[48] G. Ahdritz et al. OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. bioRxiv (2022). DOI: 10.1101/2022.11.20.517210.

[49] A. S. Lapedes, N. Santa Fe Inst., B. G. Giraud, L. C. Liu, and G. D. Stormo. Correlated mutations in protein sequences: Phylogenetic and structural effects. Lecture Notes Monogr. Ser. 33 (1999), 236–256. DOI: 10.2172/296863.

[50] D. de Juan, F. Pazos, and A. Valencia. Emerging methods in protein co-evolution. Nature Reviews Genetics 14 (4 2013), 249–261. DOI: 10.1038/nrg3414.

[51] M. Heinzinger, A. Elnaggar, Y. Wang, C. Dallago, D. Nechaev, F. Matthes, and B. Rost. Modeling aspects of the language of life through transfer-learning protein sequences. BMC Bioinformatics 20.723 (2019). DOI: 10.1186/s12859-019-3220-8.

[52] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2019.

[53] R. Wu et al. High-resolution de novo structure prediction from primary sequence. bioRxiv (2022). DOI: 10.1101/2022.07.21.500999.

[54] A. Elofsson. Progress at protein structure prediction, as seen in CASP15. Current Opinion in Structural Biology 80 (2023), 102594. DOI: https://doi.org/10.1016/j.sbi.2023.102594.

[55] S. Unsal, H. Atas, M. Albayrak, K. Turhan, A. C. Acar, and T. Doğan. Learning functional properties of proteins with language models. Nature Machine Intelligence 4 (3 2022), 227–245. DOI: 10.1038/s42256-022-00457-9.

[56] M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, and M. Steinegger. ColabFold: making protein folding accessible to all. Nature Methods 19 (6 2022), 679–682. DOI: 10.1038/s41592-022-01488-1.

[57] S. Liu et al. PSP: Million-level Protein Sequence Dataset for Protein Structure Prediction. 2022.

[58] Z. Li, X. Liu, W. Chen, F. Shen, H. Bi, G. Ke, and L. Zhang. Uni-Fold: An Open-Source Platform for Developing Protein Folding Models beyond AlphaFold. bioRxiv (2022). DOI: 10.1101/2022.08.04.502811.

[59] B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu, and Uniprot Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 31 (6 2013), 926–932. DOI: 10.1093/bioinformatics/btt473.

[60] M. Steinegger, M. Meier, M. Mirdita, H. Vöhringer, S. J. Haunsberger, and J. Söding. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics 20 (1 2019), 473. DOI: 10.1186/s12859-019-3019-7.

[61] A. Zemla. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Research 31 (13 2003), 3370–3374. DOI: 10.1093/nar/gkg571.

[62] V. Mariani, M. Biasini, A. Barbato, and T. Schwede. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics 29 (21 2013), 2722–2728. DOI: 10.1093/bioinformatics/btt473.

[63] J. Haas, A. Barbato, D. Behringer, G. Studer, S. Roth, M. Bertoni, K. Mostaguir, R. Gumienny, and T. Schwede. Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. Proteins: Structure, Function, and Bioinformatics 86 (Suppl 1 2018), 387–398. DOI: 10.1002/prot.25431.

[64] Z. Lin et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. bioRxiv (2022).

[65] A. Madani et al. Large language models generate functional protein sequences across diverse families. Nature Biotechnology (2023), 1546–1696. DOI: 10.1038/s41587-022-01618-2.

[66] A. Radford et al. Learning Transferable Visual Models From Natural Language Supervision. 2021.

[67] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-Shot Text-to-Image Generation. 2021.

[68] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. 2022.

[69] J.-B. Alayrac et al. Flamingo: a Visual Language Model for Few-Shot Learning. 2022.

[70] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. In: Proceedings of the 39th International Conference on Machine Learning. Ed. by K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato. Vol. 162. Proceedings of Machine Learning Research. 2022, 23318–23340. https://proceedings.mlr.press/v162/wang22al.html.

[71] D. Driess et al. PaLM-E: An Embodied Multimodal Language Model. 2023.

[72] R. Zellers, J. Lu, X. Lu, Y. Yu, Y. Zhao, M. Salehi, A. Kusupati, J. Hessel, A. Farhadi, and Y. Choi. MERLOT Reserve: Neural Script Knowledge Through Vision and Language and Sound. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 2022, 16375–16387.

[73] S. Reed et al. A Generalist Agent. Transactions on Machine Learning Research (2022). Featured Certification, Outstanding Certification. https://openreview.net/forum?id=1ikK0kHjvj.

[74] P. Dutta and S. Saha. Amalgamation of protein sequence, structure and textual information for improving protein-protein interaction identification. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. July 2020, 6396–6407. DOI: 10.18653/v1/2020.acl-main.570.

[75] A. Sakhovskiy and E. Tutubalina. Multimodal model with text and drug embeddings for adverse drug reaction classification. Journal of Biomedical Informatics 135 (2022), 104182. DOI: https://doi.org/10.1016/j.jbi.2022.104182.

[76] C. Edwards, T. Lai, K. Ros, G. Honke, K. Cho, and H. Ji. Translation between Molecules and Natural Language. 2022.

[77] R. Taylor, M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, and R. Stojnic. Galactica: A Large Language Model for Science. 2022.

[78] A. Tamkin, V. Liu, R. Lu, D. Fein, C. Schultz, and N. Goodman. DABS: a Domain-Agnostic Benchmark for Self-Supervised Learning. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1). 2021. https://openreview.net/forum?id=Uk2mymgn_LZ.

[79] A. Tamkin, G. Banerjee, M. Owda, V. Liu, S. Rammoorthy, and N. Goodman. DABS 2.0: Improved Datasets and Algorithms for Universal Self-Supervision. In: Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track. 2022. https://openreview.net/forum?id=ChWf1E43l4.

[80] A. Aghajanyan, L. Yu, A. Conneau, W.-N. Hsu, K. Hambardzumyan, S. Zhang, S. Roller, N. Goyal, O. Levy, and L. Zettlemoyer. Scaling Laws for Generative Mixed-Modal Language Models. 2023.

[81] K. Lu, A. Grover, P. Abbeel, and I. Mordatch. Pretrained Transformers as Universal Computation Engines. 2021.