2Bits of Protein: Efficient Protein Language Models at the Scale of 2-bits

Oliver M. Turnbull [1] **Mohamed Baioumy** [2 3] **Charlotte M. Deane** [1]

Abstract

Protein language models have become an increasingly popular tool across various biological tasks, from variant effect prediction to novel sequence generation. However, state-of-the-art models often have up to billions of parameters. Such large model architectures restrict usage to groups with the necessary compute infrastructure or necessitate the use of cloud computing, incurring substantial costs and raising data privacy concerns. In this work, we investigate a ternary protein language model, which uses low-precision weights to reduce model size, energy demand, and computational requirements, making it suitable for operation on edge devices such as laptops. This addresses privacy concerns by ensuring data remains on-device and eliminates the costs associated with cloud services. We train a ternary protein language model and benchmark it against ESM-2 (8M) using the ProteinGym benchmark, demonstrating that our model achieves comparable performance while being more suitable for edge deployment. A discussion is provided on ways to improve the ternary model to outperform ESM-2 in terms of accuracy.

1. Introduction

Protein language models (pLMs) have gained widespread popularity for their use in various biological tasks, from protein variant prediction (Cheng et al., 2023) to protein optimization (Hie et al., 2023). pLMs learn an expressive representation of protein sequences through their extensive pre-training regime on unlabelled sequences, which can be leveraged for downstream tasks.

In natural language processing (NLP), neural scaling laws (Hoffmann et al., 2022) describe the relationship between model size and performance. Similar scaling laws have been identified for pLMs (Hesslow et al., 2022), showing that increased model size often leads to increased model performance. Consequently, increasingly large pLMs have been developed, with the largest version of ESM2, a masked pLM, having 15 billion parameters (Lin et al., 2023).

However, larger models have high memory requirements and lower throughput. This poses a barrier to the use of large foundation models by academic groups and smaller companies that may not have access to the GPU clusters required to run these large models, and the high costs associated with cloud computing can be prohibitive. Additionally, the use of cloud computing also introduces privacy concerns if handling confidential and/or regulated data, a common concern in biomedical research. The energy demands of larger models also contribute to increased environmental impact, further highlighting the need for more efficient and accessible solutions. Thus, there is a need to develop more efficient pLMs that can be run on lower-resourced 'edge' devices, such as laptops. Models run on edge devices are also private by default, as the data never leaves the local device.

One approach to achieving greater efficiency is by reducing the precision of model weights, compared to the 16-bit floating points commonly used to store pLM weights. At the more extreme end, binary (Wang et al., 2023) and ternary (Ma et al., 2024) precision auto-regressive natural language models have been investigated and shown to have competitive performance with the 16-bit equivalents, while having up to a 21.7 times lower energy cost and 3.55 smaller GPU memory footprint.

In this work, we investigate training an encoder-only pLM using a ternary architecture and benchmark it against ESM-2 using ProteinGym (Notin et al., 2023). Given the limited compute available, we train an 8M parameter model. Our results show competitive performance and stable training of the ternary architecture but the model accuracy does not match ESM-2. However, based on the work of Me et al. (2024), we can expect the ternary model to outperform ESM-2 once model size is scaled up.

[1]Department of Statistics, University of Oxford [2]Robotics Institute, University of Oxford [3]Exo Labs. Correspondence to: Charlotte Deane <deane@stats.ox.ac.uk>.

Proceedings of the ICML 2024 Workshop on Accessible and Efficient Foundation Models for Biological Discovery, Vienna, Austria. 2024. Copyright 2024 by the author(s).

2. Low Precision Models

Model quantization has become an active area of research in the field of natural language models due to the increasing size of model architectures. Quantization involves reducing the precision of model weights, typically from 16-bit floating-point numbers to lower-bit representations. One popular approach is post-training quantization of models trained in 16-bit precision down to low-bit models for inference. However, this often leads to performance degradation and has not been extended below 4-bits. An alternative and promising strategy is to train models from scratch with reduced precision. BitNet, introduced by (Wang et al., 2023), showed that binary quantization {-1, 1} of transformer linear layers provided competitive performance to the 16-bit equivalent. Binary quantization is particularly appealing as it means the matrix multiplication within the linear layers consists of only addition operations, significantly decreasing energy costs. This was expanded in "BitNet b1.58" to ternary quantization {-1, 0, 1}, which increased model performance while maintaining the energy-saving benefits of binary precision (Ma et al., 2024). The introduction of a 0 value allows the model to more effectively filter out unneeded neurons. The difference in performance between the ternary and 16-bit precision model was found to decrease as the model size scaled, reaching equivalence at 3 billion parameters. For the same performance, the ternary model was 2.71 times faster, used 3.55 times less GPU memory, and used 21.7 times less energy.

To constrain linear layer weights to {-1, 0, 1} (Ma et al., 2024) used a quantization function that scales the weight matrix by its average absolute value and then rounds each value to the nearest integer among {-1, 0, 1}. The quantization function W is defined as follows:

$$\widetilde{W} = \text{RoundClip}\left(\frac{W}{\gamma + \epsilon}, -1, 1\right) \tag{1}$$

Here, the RoundClip function is defined as:

$$\text{RoundClip}(x, a, b) = \max(a, \min(b, \text{round}(x))) \tag{2}$$

The scaling factor γ is calculated as the average absolute value of the weight matrix:

$$\gamma = \frac{1}{nm} \sum_{ij} |W_{ij}| \tag{3}$$

Where: W is the weight matrix, $\widetilde{W}$ is the quantized weight matrix, γ is the scaling factor, ϵ is a small constant to prevent division by zero, round(x) rounds x to the nearest integer, a and b are the clipping boundaries, set to -1 and 1 respectively.

The rest of the architecture is identical to the regular transformer encoder architecture used in ESM2 (Lin et al., 2023). The linear layers in the encoder are simply replaced by the ternary quantized version defined above.

3. Methods

3.1. Model Architecture

For our initial study, we modified the smallest available ESM-2 model with 8 million parameters (Lin et al., 2023). We followed the methodology of Ma et al. (2024) by implementing ternary precision of weights in all linear layers, apart from the embedding and language modeling head layers. We used the same Uniref50 (Suzek et al., 2015) dataset and train/validation splits as ESM2.

3.2. Training

Following the work of Frey et al. (2024), we trained a 'crammed' model, in which training is restricted to 24 GPU hours, allowing for quick iteration and benchmarking. The crammed model uses a reduced context length of 512 which allows for a much larger batch size. For hyper-parameter tuning, we used a random search around the parameters used by Frey et al. (2024), see Appendix A.1 for the hyper-parameters used to train "Crammed Ternary ESM2 (8M)". As found by Ma et al. (2024), we saw stability at higher learning rates for the 1-bit architecture compared to mixed precision models, with an optimal learning rate of 5E-3. We observed the reported S-shaped loss curve with a sharp decrease in loss towards the end of training (Figure 1).

We also performed initial training of a ternary model for the same number of epochs as ESM-2 (8M) ("Ternary ESM2 (8M)"), but, we were only able to train for 6 out of 15 epochs due to time and compute constraints.

All training was performed on NVIDIA A100 80GB GPUs using PyTorch and HuggingFace Transformers.

4. Results

We evaluated our ternary model using ProteinGym, which provides a large set of benchmarks for protein fitness prediction. We used the 217 deep-mutational scan datasets for zero-shot fitness prediction across various protein classes and functional tasks, providing a rigorous assessment of model performance. For our baseline, we used the ESM2 (8M) results provided by ProteinGym. We also trained and evaluated a "Crammed ESM2 (8M)" model using the hyper-parameters and procedure described in (Frey et al., 2024) with the standard ESM2 architecture.

Table 1 shows the average Spearman correlation across DMS datasets for the ternary and baseline models. Ternary

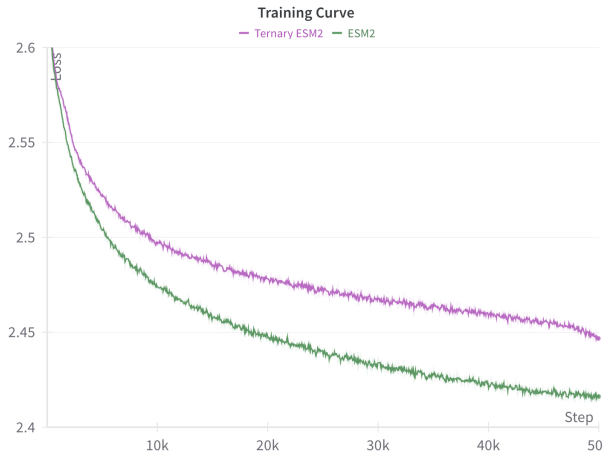

Figure 1. Training loss curve of the crammed baseline ESM2 (8M) model and ternary ESM2. The ternary model shows a characteristic "S-shaped" loss curve with a sudden decrease in loss towards the end of training.

Model	Avg. Spearman	SD
ESM2-8M	0.226	0.019
Crammed ESM2-8M	0.192	0.019
Ternary ESM2-8M	0.181	0.019
Crammed Ternary ESM2-8M	0.166	0.020

Table 1. Average Spearman Correlation on ProteinGym Zero-Shot Tasks. ESM2-8M values were obtained from the results submitted to ProteinGym. Crammed ESM2-8M and Ternary ESM2-8M values were obtained from our experiments.

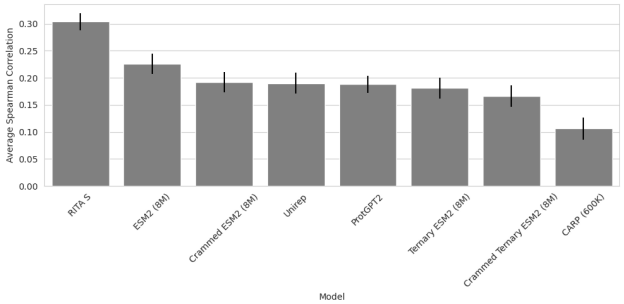

Figure 2. Spearman correlation for a selection of pLMs on the ProteinGym Zero-Shot benchmark. Error bars represent the standard deviation of the mean Spearman correlation, estimated by bootstrapping with 10,000 samples.

ESM2 achieves an average Spearman correlation of 0.181, making it competitive with ProtGPT2 and UniRep (Figure 2). The baseline ESM2 (8M) model had a 25% higher Spearman correlation than the Ternary ESM2 model, which was found to be statistically significant at the 5% confidence level (see Appendix A.2 for details). However, the baseline model also had roughly double the compute budget of the ternary model. The crammed models provide a fairer test, as both the ternary and baseline model had the same compute budget. Here, the baseline "Crammed ESM2" only outperformed the "Crammed Ternary ESM2" by 15% and furthermore, this difference was not statistically significant.

If pLMs follow the same trend as observed in NLP LMs, we would expect the performance difference between ternary and 16-bit models to decrease as model size scales, as well as an increasing model compression ratio. It would be interesting to see if this holds for pLMs too.

Even if the performance of ternary models are not able to fully match that of 16-bit models, given the performance we have already shown they could still serve as a valuable tool for pre-screening sequences. The significantly reduced inference cost and energy usage of ternary models make them well-suited for initial screening of large databases of sequences. By using a ternary model to pre-filter sequences before passing them through a more computationally expensive 16-bit precision model, the overall computational burden and associated environmental and financial costs can be significantly reduced.

5. Conclusion

We presented initial work on applying ternary precision to pLMs. We showed that a ternary precision implementation of ESM2 achieves competitive performance with the baseline model, despite the small architecture size. This highlights the opportunities of applying ternary precision to pLMs, providing significant reductions in inference costs and environmental impact. Given the work of Ma et al. (2024), we would expect the performance of the ternary precision model to scale with model size and match that of the full precision ESM-2 while having faster inference time, significantly lower energy consumption and lower memory requirements. This also opens up the opportunity to apply low-precision techniques to other costly biological models such as inverse folding (Hsu et al., 2022) and structure prediction models (Abramson et al., 2024) to deliver performance gains to these areas as well.

Acknowledgements

This work was supported by funding from the UK Engineering and Physical Sciences Research Council (OT, reference EP/S024093/1) and AstraZeneca.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., Beattie, C., Bertolli, O., Bridgland, A., Cherepanov, A., Congreve, M., Cowen-Rivers, A. I., Cowie, A., Figurnov, M., Fuchs, F. B., Gladman, H., Jain, R., Khan, Y. A., Low, C. M. R., Perlin, K., Potapenko, A., Savy, P., Singh, S., Stecula, A., Thillaisundaram, A., Tong, C., Yakneen, S., Zhong, E. D., Zielinski, M., Žídek, A., Bapst, V., Kohli, P., Jaderberg, M., Hassabis, D., and Jumper, J. M. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, pp. 1–3, May 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07487-w. URL https://www.nature.com/articles/s41586-024-07487-w. Publisher: Nature Publishing Group.

Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L. H., Zielinski, M., Sargeant, T., Schneider, R. G., Senior, A. W., Jumper, J., Hassabis, D., Kohli, P., and Avsec, Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, 381 (6664):eadg7492, September 2023. doi: 10.1126/science.adg7492. URL https://www.science.org/doi/10.1126/science.adg7492. Publisher: American Association for the Advancement of Science.

Frey, N. C., Joren, T., Ismail, A. A., Goodman, A., Bonneau, R., Cho, K., and Gligorijević, V. Cramming Protein Language Model Training in 24 GPU Hours. *bioRxiv*, pp. 2024.05.14.594108, January 2024. doi: 10.1101/2024.05.14.594108. URL http://biorxiv.org/content/early/2024/05/15/2024.05.14.594108.abstract.

Hesslow, D., Zanichelli, N., Notin, P., Poli, I., and Marks, D. RITA: a Study on Scaling Up Generative Protein Sequence Models, July 2022. URL http://arxiv.org/abs/2205.05789. arXiv:2205.05789 [cs, q-bio].

Hie, B. L., Shanker, V. R., Xu, D., Bruun, T. U. J., Weidenbacher, P. A., Tang, S., Wu, W., Pak, J. E., and Kim, P. S. Efficient evolution of human antibodies from general protein language models. *Nature Biotechnology*, pp. 1–9, April 2023. ISSN 1546-1696. doi: 10.1038/s41587-023-01763-2. URL https://www.nature.com/articles/s41587-023-01763-2. Publisher: Nature Publishing Group.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Driessche, G. v. d., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. Training Compute-Optimal Large Language Models, March 2022. URL http://arxiv.org/abs/2203.15556. arXiv:2203.15556 [cs].

Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A., and Rives, A. Learning inverse folding from millions of predicted structures, September 2022. URL https://www.biorxiv.org/content/10.1101/2022.04.10.487779v2. Pages: 2022.04.10.487779 Section: New Results.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, March 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/10.1126/science.ade2574. Publisher: American Association for the Advancement of Science.

Ma, S., Wang, H., Ma, L., Wang, L., Wang, W., Huang, S., Dong, L., Wang, R., Xue, J., and Wei, F. The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits, February 2024. URL http://arxiv.org/abs/2402.17764. arXiv:2402.17764 [cs] version: 1.

Notin, P., Kollasch, A., Ritter, D., van Niekerk, L., Paul, S., Spinner, H., Rollins, N., Shaw, A., Orenbuch, R., Weitzman, R., Frazer, J., Dias, M., Franceschi, D., Gal, Y., and Marks, D. ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 64331–64379. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/cac723e5ff29f65e3fcbb0739ae91bee-Paper-Datasets_and_Benchmarks.pdf.

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., and the UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, March 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu739. URL https://doi.org/10.1093/bioinformatics/btu739.

Wang, H., Ma, S., Dong, L., Huang, S., Wang, H., Ma, L., Yang, F., Wang, R., Wu, Y., and Wei, F. BitNet: Scaling 1-bit Transformers for Large Language Models, Oc-

tober 2023. URL http://arxiv.org/abs/2310.11453. arXiv:2310.11453 [cs].

A. Appendix

A.1. Model Hyperparameters

All models were trained using a weight decay of 0.01, bf16 mixed precision, and the AdamW optimizer. The crammed models were able to use a much larger batch size due to the decreased memory requirements from the lower context length.

Hyperparameter	Crammed ESM	Crammed Ternary ESM2	Ternary ESM2
Warmup steps	1,000	1,000	20,000
Total steps	50,000	50,000	15 epochs
Learning Rate	1E-3	5E-3	4E-3
Batch size	256	256	32
Gradient accumulation steps	8	8	32
Context Length	512	512	1,024

Table 2. Model hyperparameters for Crammed ESM, Crammed Ternary ESM2, and Ternary ESM2.

A.2. Statistical Testing

Table 3 shows the results of the one-tailed Z-tests comparing the mean Spearman correlation of ESM2 vs Ternary ESM2 on the ProteinGym DMS benchmark. The alternative hypothesis (H1) for each test states that the mean Spearman correlation of the baseline ESM2 model is greater than the Ternary ESM2 model. The mean Spearman correlation standard deviation for each model was estimated using bootstrapping with 10,000 samples, as implemented in ProteinGym. The results show that the ESM2 (8M) model's performance is significantly better than that of the Ternary ESM2 (8M) model at the 5% level. In comparison, there is no significant difference between the Crammed ESM2 (8M) and Crammed Ternary ESM2 (8M) models.

Alternative Hypothesis (H1)	Z-score	p-value	Significant at 5% level
$\mu_{ESM2} > \mu_{Ternary}$	1.67	0.047	Yes
$\mu_{CrammedESM2} > \mu_{CrammedTernary}$	0.94	0.17	No

Table 3. Results of one-tailed Z-tests comparing the mean Spearman correlations (μ) of ESM2 (8M) vs. Ternary ESM2 (8M) and Crammed ESM2 (8M) vs. Crammed Ternary ESM2 (8M) on the ProteinGym benchmark.