ASSESSING QUANTIZATION AND EFFICIENT FINE-TUNING FOR PROTEIN LANGUAGE MODELS

Sebastian Clancy[1*] **Ilan Yaniv Zeisler**[1*] **Pouriya Bayat**[1] **Matthew Xie**[2]

Vivian White[2] **Spencer Perkins**[2] **Sepehr Bayat**[3] **Keith Pardee**[1]

[*]These authors contributed equally
[1]Leslie Dan Faculty of Pharmacy, University of Toronto
[2]Department of Computer Science, University of Toronto
[3]Computing and Software, McMaster University

ABSTRACT

Proteins are essential to life and function, and discovering new proteins can unlock new therapeutics and industrial applications. However, the space of proteins is incredibly large and diverse, making discovering useful proteins difficult. Machine learning (ML) models help search the space of proteins, finding candidate proteins for specific goals and reducing the need for costly experimentation. The recent trend of increasing scale of ML models creates more demanding computational requirements, especially for large language models (LLMs) and their protein language model (PLM) counterparts. Quantization and efficient fine-tuning methods can help offset this by reducing the amount of memory and training required to use ML models. Here we show that combining 4-bit quantization and efficient training with low rank adapters maintains $>$90% of the performance for most models in protein prediction tasks, while simultaneously reducing the required memory consumption by 46.7% on average. Generative models that are 4-bit quantized use 76.4% less memory while showing no significant difference in the quality of their generated proteins. This represents the first benchmark of quantized training with parameter efficient fine-tuning for PLMs while retaining nearly all of their performance, thus lowering the requirements and barrier of entry for practitioners.

1 INTRODUCTION

Proteins represent a highly versatile class of molecules with applications in medicine, biotechnology and chemistry. However, the immense space of possible proteins (20^{100} possible sequences for a 100 amino acid protein) presents a very challenging problem in discovering proteins of interest. This is compounded by the associated high cost and time of testing these proteins in the laboratory. ML models are capable of recognizing underlying patterns in data (Frappier & Keating, 2021) and have recently become remarkably successful in the design of new proteins (Hayes et al., 2025; Watson et al., 2023). Protein language models are large language models trained on protein amino acid sequences instead of natural language, and are often fine-tuned for specific downstream tasks such as enzymatic reaction performance (Chen et al., 2025) or drug binding kinetics (Paul et al., 2021).

ML models and LLMs/PLMs in particular have heavy computational requirements which have only increased with the recent trend in the field to use massively larger models (see Figure 1). Some models have even reached trillions of model parameters, and PLMs are following this trend as well (Zhao et al., 2023; Vieira et al., 2025). As PLMs scale in size, so do the GPU memory requirements resulting in a higher training/inference cost and barrier to usage, particularly for academic labs and small companies. Methods that reduce PLM compute requirements are key to mitigating these ballooning costs.

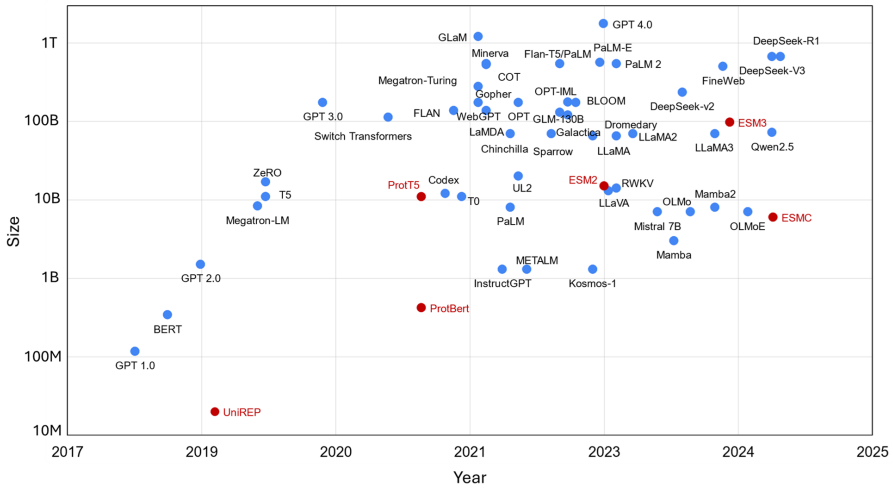

Figure 1: Parameter count of landmark LLMs (blue) and PLMs (red) over time (Amatriain et al., 2023). For models with multiple sizes, the model with highest performance was selected. GPT 4.0 model size is a community estimate given lack of published statistics.

Current methods for reducing the computational cost of large models focus either on creating new architectures that use fewer parameters, or modifying existing training and inference methods to reduce size (Han et al., 2024). We focus on methods that integrate with existing models, in particular quantization and parameter efficient fine tuning (PEFT) methods. Low Rank Adaptation (LoRA) is a PEFT method that freezes the existing model weights and substitutes rank decomposed matrices to train instead. Hu et al. (2021) demonstrated that the modifications in weight matrices during fine-tuning are typically low-rank, which permits the approximation of these changes through a decomposition into two smaller matrices. These substitute weights greatly reduce the trainable parameters while retaining a comparable level of detail in the model during fine-tuning.

Schmirler et al. (2024) showed that PLMs like ESM-2 and ProtT5 can be efficiently fine-tuned with LoRA to improve protein prediction and reduce computational costs. Sledzieski et al. (2024) further applied LoRA to proteomics, demonstrating its effectiveness in protein-protein interaction (PPI) prediction and homo-oligomer symmetry classification. These methods can be extended for further efficiency benefits by combining quantization with LoRA (QLoRA), as has been done with LLMs (Dettmers et al., 2024). Our work builds on these efforts by systematically applying QLoRA across diverse protein modeling tasks, thus optimizing memory efficiency while maintaining high performance across various models.

Here, we apply QLoRA for fine-tuning PLMs across multiple model scales and designs, including ESM-2, ESM C, ProtBERT, ProtT5-half, and Ankh-base. We quantify model performance on both regression (predicting fluorescent protein brightness and protease stability) and classification (predicting secondary structure) tasks. By comparing QLoRA-based approaches against traditional full-model fine-tuning, we highlight the potential to reduce memory footprints while maintaining strong predictive performance. We also apply quantization to ProtGPT2 and ProLLaMA to determine the effects on unconditional generation of proteins. Our experiments confirm the practical feasibility of applying PEFT and quantization methods together for large-scale protein modeling, opening up avenues for more accessible and democratized protein design processes.

2 METHODS

2.1 PRE-TRAINED PLMS

All models are based on the transformer architecture and initialized with pre-trained checkpoints from the Hugging Face Transformers library. ESM-2 models ranging from 8 million to 3 billion parameters (8M, 35M, 150M, 650M, 3B), ESM C (300M, 600M), ProtBERT (420M), ProtT5-half

(3B), and Ankh-Base (726M) were used for fine-tuning with QLoRA. ProtGPT2 (738M) and ProLLaMA (6.7B) were used for unconditional generation with quantization.

2.2 Compute Benchmark

We evaluated several metrics of computational cost for training each model. We tracked the maximum GPU memory utilized by the model and the average power consumption of the model during training. For each metric, we trained each model for various batch sizes ($2^n, n \in 0, 1, \ldots, 8$) and measured the metric with the full model, and again with the quantized model. We evaluated these on a single dataset, green fluorescent protein (GFP) brightness prediction, with three seeded replicates. We tracked model consumption metrics with the wandb library and nvidia-stat-gpu.

2.3 QLoRA Fine-tuning

We used the following datasets for fine-tuning: For regression, we used GFP brightness (Sarkisyan et al., 2016) and protease stability (Rocklin et al., 2017). For 3-class classification, we used protein secondary structure (Berman et al., 2000; Moult et al., 2018; Klausen et al., 2019).

We performed an initial hyperparameter search on ESM-2 150M across learning rate (LR), alpha, and rank, and used this as a basis for fine-tuning all the models with adjustments to LR and batch size to facilitate convergence as needed (Appendix A). For QLoRA models, we quantized to 4-bit float on linear layers in the base model (either by default layers specified by the model implementation in the Transformers library or our specification if there was no default) and used LoRA rank=8, alpha=32 and dropout=0.05. We fine-tuned each model on each dataset task for both the full model and the QLoRA model in triplicate across three seeds for 10 epochs. We evaluated models with Spearman's rank correlation coefficient (SpearmanR) for fluorescence and stability regression tasks and multi-class accuracy for the secondary structure classification task.

2.4 Quantized Unconditional Generation

We evaluated full and 4-bit quantized models in triplicate across three seeds each by generating 100 sequences between 100-200 amino acids and screening the generated sequences for predicted structure based scores. We quantized ProLLaMA on the `q_proj`, `k_proj`, `v_proj` and `o_proj` layers, and ProtGPT2 on the `c_attn` and `c_proj` layers. We used Foldseek (Van Kempen et al., 2024) to evaluate pass rate (ratio of sequences with any homology match), local distance difference test (LDDT), alignment score and homology probability. We used s4pred (Moffat & Jones, 2021) to determine helix, sheet and coil content. We used aiupred (Erdős & Dosztányi, 2024) to determine disorder probability. Lastly, we used AlphaFold2 (Jumper et al., 2021) via ColabFold (Mirdita et al., 2022) to determine predicted LDDT (pLDDT) and predicted template modeling (pTM) scores.

3 Results

3.1 Compute Performance

All models saw reduced memory usage with small batch sizes and tended towards a larger reduction for larger models (Figure 2, Appendix B.2). However, as batch size increased, we saw less of a reduction, even going as far as increasing memory size for the smallest models. For the largest models, quantization reduced the memory to just 10% of their original size. The average memory reduction across all batch sizes and models was 46.7%. All models showed reduced power consumption on average (Appendix B.1). Overall, for all models and batch sizes we saw an average 5.6% less power consumed.

3.2 Fine-Tuning Performance

We fine-tuned each model on each dataset with QLoRA models performing generally close to the full model performance for all models (Figure 3A). For smaller models, batch size and learning rate could be kept the same for full and QLoRA models to obtain convergence, while larger models tended to require a magnitude lower learning rate for the full model relative to the QLoRA model

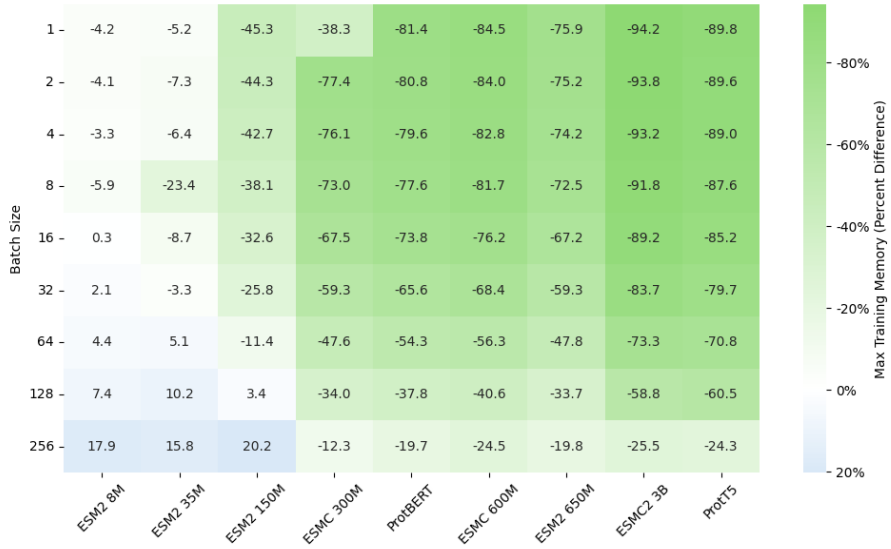

Figure 2: Heatmap showing the difference in model training memory across models and batch sizes. We tested each model in triplicate on a 1xH100 GPU. Percent difference is calculated by $\frac{full-qlora}{full} \times 100\%$. Negative numbers represent reduced memory consumption, positive increased.

(Appendix A). ESM C 600M performed best in both overall fine-tuning performance and relative QLoRA vs. full model performance with fluorescence SpearmanR 0.850 vs. 0.863, stability SpearmanR 0.556 vs. 0.671 and secondary structure accuracy 0.870 vs. 0.870. All models had relatively poor performance on the stability dataset compared to fluorescence and secondary structure. The ESM-2 family exhibited better relative QLoRA performance as parameterization increased, and larger models also generally exhibited the same better performance relative to smaller models. ESM C 300M and 600M showed performance similar or better than the much larger ESM-2 3B. The T5 based models, ProtT5 and Ankh, were both challenging to train and exhibited high sensitivity to LR and batch size with large variance.

3.3 Unconditional Generation Quality

We found ProtGPT2 and ProLLaMA both quantized well with very similar predicted structural properties of the generated proteins (Figure 3B). Through Foldseek, full and quantized models exhibited very similar characteristics though ProtGPT2 and ProLLaMA had lower pass rates for quantized vs. full of 84.0% vs. 89.3% and 83.3% vs. 89.3%, respectively. Secondary structures predicted by s4pred were also similar between model types with preference towards disordered structures. Overall disorder predicted by aiupred was lower for quantized ProtGPT2 (36.5% quantized vs. 40.2% full) and higher for quantized ProLLaMA (54.2% quantized vs. 50.9% full). AlphaFold2 predicted close pLDDT and pTM for quantized and full versions of both models. Overall, both models were not significantly different in their generated output characteristics between full and quantized models (Welch's t-test, ProtGPT2 p=0.805, ProLlaMa p=0.787).

4 Discussion

We have shown that PLMs can be quantized and efficiently trained without sacrificing significant performance. Fine-tuning across various tasks with 4-bit conversion of the large majority of model weights and utilization of LoRA provides much smaller memory footprints with mostly preserved performance in these contexts. We saw QLoRA performance get closer to full model parity with lower deviation across seeds as the model size increased, both within the ESM-2 family and for all models in general. ESM C 600M in particular quantized very well and performed better than its much larger but earlier generation counterpart ESM-2 3B, in line with ESM Team (2024). The

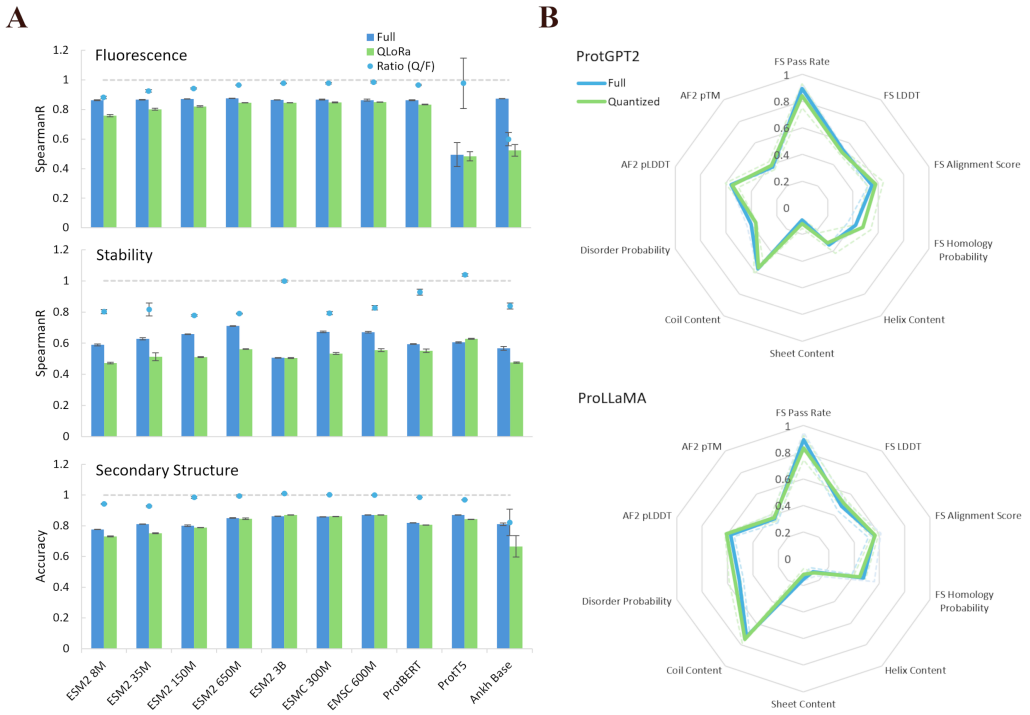

Figure 3: PLMs can be efficiently fine-tuned and quantized for prediction and generation. (A) Comparison of full and QLoRA models after fine-tuning across each dataset. Dots above bars represent relative performance of quantized to full. Error bars represent standard deviation across seeds. (B) Comparison of full and 4-bit quantized PLMs for unconditional generation of 100-200 length proteins. FS pass rate, FS LDDT, FS alignment score, FS homology probability are determined from FoldSeek. Helix, sheet and coil content are determined from s4pred. Disorder probability is determined from aiupred. pTM and pLDDT are determined with AlphaFold2. Dashed lines represent 95% CI upper and lower bounds across seeds.

unconditional generation of new sequences is also preserved for both ProtGPT2 and ProLLaMA when 4-bit quantized, allowing nearly identical quality proteins with a vast reduction in memory.

QLoRA reduced the memory utilization of all models at small batch sizes, however this reduction was lost for large batch sizes, even increasing memory requirements in some instances. This is line with Dettmers et al. (2024), where increased batch sizes results in a larger memory footprint from the activation gradient. For the models and batch sizes that see a large reduction in memory, it is much easier to run these on readily available hardware. The larger models we tested (ProtT5, ESM-2 3B) initially took up $\sim$60 GB, too large to fit on an RTX A6000 GPU. After quantization, these models were reduced to less than 10 GB, able to train and run on a free Colaboratory notebook with a T4 GPU.

Our work demonstrates a promising application of QLoRA to PLMs, showing computational benefits with minimal performance trade-offs. Achieving a 90% reduction in training memory while retaining at least 90% of performance represents a significant advantage for practical model training and inference. While model memory requirements increased with larger batch sizes, forthcoming GPU architectures with native support for sub 8-bit floating point operations may help address this limitation (NVIDIA, 2024). The ability to achieve similar performance with more efficient computation enables both a wider exploration of protein space and accessibility for researchers in the field.

REFERENCES

Xavier Amatriain, Ananth Sankar, Jie Bing, Praveen Kumar Bodigutla, Timothy J Hazen, and Michaeel Kazi. Transformer models: an introduction and catalog. *arXiv preprint arXiv:2302.07730*, 2023.

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1): 235–242, 2000.

Jia-Ying Chen, Jing-Fu Wang, Yue Hu, Xin-Hui Li, Yu-Rong Qian, and Chao-Lin Song. Evaluating the advancements in protein language models for encoding strategies in protein function prediction: a comprehensive review. *Frontiers in Bioengineering and Biotechnology*, 13:1506508, 2025.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024.

Gábor Erdős and Zsuzsanna Dosztányi. Aiupred: combining energy estimation with deep learning for the enhanced prediction of protein disorder. *Nucleic Acids Research*, pp. gkae385, 2024.

ESM Team. Esm cambrian: Revealing the mysteries of proteins with unsupervised learning, 2024. URL https://evolutionaryscale.ai/blog/esm-cambrian.

Vincent Frappier and Amy E Keating. Data-driven computational protein design. *Current Opinion in Structural Biology*, 69:63–69, 2021.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey, 2024. URL https://arxiv.org/abs/2403.14608.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, pp. eads0018, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Michael Schantz Klausen, Martin Closter Jespersen, Henrik Nielsen, Kamilla Kjaergaard Jensen, Vanessa Isabell Jurtz, Casper Kaae Soenderby, Morten Otto Alexander Sommer, Ole Winther, Morten Nielsen, Bent Petersen, et al. Netsurfp-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins: Structure, Function, and Bioinformatics*, 2019.

Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. Colabfold: making protein folding accessible to all. *Nature methods*, 19(6):679–682, 2022.

Lewis Moffat and David T Jones. Increasing the accuracy of single sequence prediction methods using a deep semi-supervised learning framework. *Bioinformatics*, 37(21):3744–3751, 2021.

John Moult, Krzysztof Fidelis, Andriy Kryshtafovych, Torsten Schwede, and Anna Tramontano. Critical assessment of methods of protein structure prediction (CASP)-Round XII. *Proteins: Structure, Function, and Bioinformatics*, 86:7–15, 2018. ISSN 08873585. doi: 10.1002/prot.25415. URL http://doi.wiley.com/10.1002/prot.25415.

NVIDIA. NVIDIA Blackwell Datasheet, 2024. URL https://resources.nvidia.com/en-us-blackwell-architecture/datasheet.

Debleena Paul, Gaurav Sanap, Snehal Shenoy, Dnyaneshwar Kalyane, Kiran Kalia, and Rakesh K. Tekade. Artificial intelligence in drug discovery and development. *Drug Discovery Today*, 26(1): 80–93, 2021. ISSN 1359-6446. doi: https://doi.org/10.1016/j.drudis.2020.10.010. URL https://www.sciencedirect.com/science/article/pii/S1359644620304256.

Gabriel J Rocklin, Tamuka M Chidyausiku, Inna Goreshnik, Alex Ford, Scott Houliston, Alexander Lemak, Lauren Carter, Rashmi Ravichandran, Vikram K Mulligan, Aaron Chevalier, et al. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science*, 357 (6347):168–175, 2017.

Karen S Sarkisyan, Dmitry A Bolotin, Margarita V Meer, Dinara R Usmanova, Alexander S Mishin, George V Sharonov, Dmitry N Ivankov, Nina G Bozhanova, Mikhail S Baranov, Onuralp Soylemez, et al. Local fitness landscape of the green fluorescent protein. *Nature*, 533(7603):397–401, 2016.

Robert Schmirler, Michael Heinzinger, and Burkhard Rost. Fine-tuning protein language models boosts predictions across diverse tasks. *Nature Communications*, 15:7407, 2024. doi: 10.1038/s41467-024-51844-2.

Samuel Sledzieski, Meghana Kshirsagar, Minkyung Baek, Rahul Dodhia, Juan Lavista Ferres, and Bonnie Berger. Democratizing protein language models with parameter-efficient fine-tuning. *Proceedings of the National Academy of Sciences (PNAS)*, 121(26):e2405840121, 2024. doi: 10.1073/pnas.2405840121.

Michel Van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature biotechnology*, 42(2):243–246, 2024.

Luiz C. Vieira, Morgan L. Handojo, and Claus O. Wilke. Scaling down for efficiency: Medium-sized protein language models perform well at transfer learning on realistic datasets. *bioRxiv*, 2025. doi: 10.1101/2024.11.22.624936. URL https://www.biorxiv.org/content/early/2025/01/28/2024.11.22.624936.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

A Extended Methods

	Fluorescence				Stability				Secondary Structure			
	Full		QLoRA		Full		QLoRA		Full		QLoRA	
Model	BS	LR	BS	LR	BS	LR	BS	LR	BS	LR	BS	LR
ESM-2 8M	32	5E-05	32	5E-05	32	5E-05	32	5E-05	32	5E-05	32	5E-05
ESM-2 35M	32	5E-05	32	5E-05	32	5E-05	32	5E-05	32	5E-05	32	5E-05
ESM-2 150M	32	5E-05	32	5E-05	32	5E-05	32	5E-05	32	5E-06	32	5E-05
ESM-2 650M	16	5E-05	16	5E-05	16	5E-05	16	5E-05	16	5E-06	16	5E-05
ESM-2 3B	32	5E-06	32	5E-05	32	5E-06	32	5E-05	8	5E-06	4	5E-05
ESM C 300M	32	5E-05	32	5E-05	32	5E-05	32	5E-05	32	5E-05	32	5E-05
ESM C 600M	32	5E-05	32	5E-05	32	5E-05	32	5E-05	16	5E-06	32	5E-05
ProtBERT	32	5E-06	32	5E-05	32	5E-06	32	5E-05	32	5E-06	32	5E-05
ProtT5 Half	8	5E-08	8	5E-07	32	5E-06	32	5E-05	4	5E-06	16	5E-05
Ankh Base	8	5E-05	32	5E-06	8	5E-05	8	5E-05	8	5E-05	8	5E-05

Table 1: Batch size and learning rate for full and QLoRA models fine-tuned on fluorescence, stability and secondary structure datasets. LoRA parameters were fixed with $r = 8, \alpha = 32, dropout = 0.05$. Runs were completed on NVIDIA GeForce RTX 4090 24GB, RTX A6000 48GB and H100 80GB depending on model size and availability.

B Compute Benchmarks

B.1 Power Consumption Difference

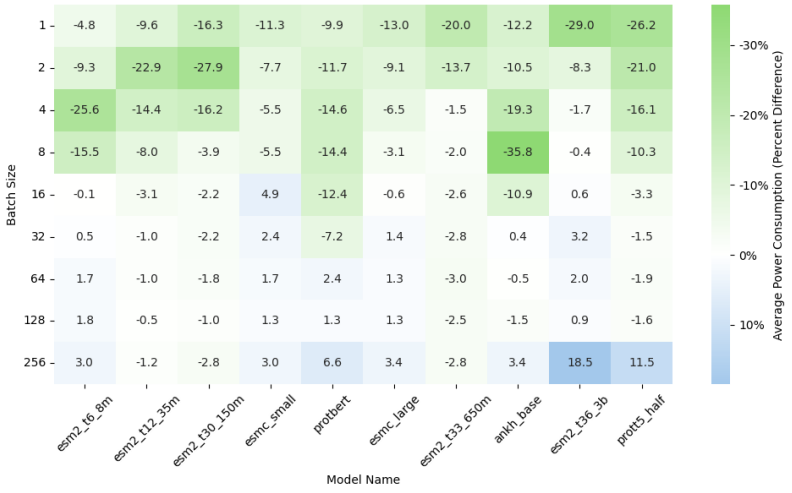

Figure 4: Heatmap showing percent difference in power consumption for models. Tests were ran in triplicate across a 1xH100 cluster, measuring average GPU power consumption during training.

B.2 RAW COMPUTE METRICS

ESM2 8M

	Full		QLoRA	
Batch Size	**Memory (MB)**	**Power (W)**	**Memory (MB)**	**Power (W)**
1	$3,633.6 \pm 1.4$	91.6 ± 3.3	$3,480.5 \pm 1.4$	87.2 ± 0.9
2	$3,655.9 \pm 0.0$	108.1 ± 1.1	$3,507.0 \pm 0.0$	98.0 ± 0.8
4	$3,699.9 \pm 0.0$	157.2 ± 1.3	$3,578.3 \pm 0.0$	117.0 ± 0.3
8	$3,957.9 \pm 0.0$	219.1 ± 1.9	$3,725.1 \pm 0.0$	185.2 ± 0.2
16	$3,980.9 \pm 0.0$	239.3 ± 1.4	$3,993.5 \pm 0.0$	239.0 ± 1.0
32	$4,429.7 \pm 0.0$	242.8 ± 0.8	$4,522.0 \pm 0.0$	244.0 ± 0.9
64	$5,321.0 \pm 0.0$	241.2 ± 0.2	$5,553.8 \pm 0.0$	245.2 ± 0.6
128	$7,115.7 \pm 4.5$	241.4 ± 2.1	$7,643.4 \pm 2.7$	245.7 ± 1.4
256	$12,873.8 \pm 1,864.9$	237.5 ± 2.3	$15,177.1 \pm 231.4$	244.5 ± 1.3

ESM2 35M

	Full		QLoRA	
Batch Size	**Memory (MB)**	**Power (W)**	**Memory (MB)**	**Power (W)**
1	$8,328.9 \pm 397.9$	103.5 ± 1.7	$7,899.1 \pm 64.2$	93.5 ± 0.3
2	$9,104.9 \pm 35.0$	136.7 ± 3.8	$8,440.5 \pm 37.8$	105.5 ± 1.6
4	$9,136.1 \pm 73.4$	185.9 ± 3.9	$8,551.7 \pm 85.2$	159.2 ± 2.9
8	$7,040.9 \pm 2,194.7$	252.0 ± 0.7	$5,394.3 \pm 2,362.5$	231.7 ± 2.1
16	$4,803.7 \pm 5.3$	265.0 ± 1.0	$4,386.8 \pm 3.4$	256.9 ± 0.9
32	$5,239.1 \pm 0.0$	266.7 ± 1.1	$5,065.1 \pm 0.0$	264.2 ± 6.1
64	$6,159.8 \pm 0.0$	266.5 ± 0.7	$6,472.3 \pm 0.0$	263.8 ± 0.4
128	$8,311.5 \pm 0.0$	266.2 ± 0.3	$9,158.7 \pm 0.0$	264.8 ± 0.5
256	$12,621.1 \pm 0.0$	265.8 ± 0.1	$14,614.1 \pm 1.2$	262.7 ± 0.5

ESM2 150M

	Full		QLoRA	
Batch Size	**Memory (MB)**	**Power (W)**	**Memory (MB)**	**Power (W)**
1	$6,949.7 \pm 1.2$	118.7 ± 0.9	$3,800.4 \pm 0.0$	99.4 ± 0.4
2	$6,935.6 \pm 0.1$	162.3 ± 2.1	$3,863.2 \pm 0.0$	117.1 ± 0.8
4	$6,931.3 \pm 0.0$	224.0 ± 2.9	$3,972.3 \pm 0.0$	187.7 ± 1.7
8	$6,950.2 \pm 0.0$	260.5 ± 0.8	$4,299.4 \pm 0.0$	250.4 ± 0.8
16	$7,147.4 \pm 0.0$	266.0 ± 1.0	$4,819.5 \pm 0.0$	260.2 ± 0.2
32	$7,931.7 \pm 0.0$	270.0 ± 1.2	$5,889.1 \pm 0.0$	264.0 ± 0.7
64	$9,162.7 \pm 0.0$	269.1 ± 0.4	$8,122.5 \pm 0.0$	264.2 ± 0.5
128	$12,094.5 \pm 0.0$	268.3 ± 0.6	$12,509.8 \pm 0.0$	265.7 ± 0.3
256	$17,670.9 \pm 0.0$	269.3 ± 0.9	$21,248.6 \pm 0.0$	261.9 ± 1.6

ESM2 650M

	Full		QLoRA	
Batch Size	**Memory (MB)**	**Power (W)**	**Memory (MB)**	**Power (W)**
1	$17,389.8 \pm 0.0$	209.8 ± 2.8	$4,184.1 \pm 0.0$	167.8 ± 1.0
2	$17,429.7 \pm 0.0$	251.5 ± 1.7	$4,320.4 \pm 0.0$	217.0 ± 3.1
4	$17,706.5 \pm 0.0$	264.5 ± 1.4	$4,565.8 \pm 0.0$	260.7 ± 0.5
8	$17,859.6 \pm 0.0$	269.5 ± 0.6	$4,907.6 \pm 0.0$	264.1 ± 0.0
16	$17,566.0 \pm 0.0$	272.6 ± 0.2	$5,765.3 \pm 0.0$	265.5 ± 0.6
32	$18,199.3 \pm 0.0$	273.7 ± 0.6	$7,415.8 \pm 0.0$	266.0 ± 0.2
64	$20,571.2 \pm 0.0$	273.7 ± 0.2	$10,746.1 \pm 0.0$	265.5 ± 0.2
128	$26,244.0 \pm 0.0$	273.0 ± 0.4	$17,391.9 \pm 0.0$	266.1 ± 0.4
256	$38,283.8 \pm 0.0$	269.1 ± 0.2	$30,702.6 \pm 0.0$	261.5 ± 0.9

ESM2 3B

	Full		QLoRA	
Batch Size	**Memory (MB)**	**Power (W)**	**Memory (MB)**	**Power (W)**
1	$61,072.3 \pm 0.0$	227.4 ± 19.9	$3,511.8 \pm 0.0$	161.5 ± 15.9
2	$61,393.2 \pm 0.0$	234.1 ± 15.6	$3,782.3 \pm 0.0$	214.7 ± 31.0
4	$61,521.1 \pm 0.0$	257.9 ± 0.8	$4,203.9 \pm 0.0$	253.5 ± 3.1
8	$62,414.5 \pm 0.0$	260.0 ± 4.4	$5,091.0 \pm 0.0$	258.8 ± 5.7
16	$61,370.1 \pm 0.0$	259.8 ± 3.9	$6,615.6 \pm 0.0$	261.4 ± 5.4
32	$61,972.0 \pm 0.0$	252.8 ± 2.4	$10,086.4 \pm 0.0$	260.9 ± 4.6
64	$64,379.6 \pm 0.0$	255.2 ± 5.1	$17,168.5 \pm 0.0$	260.3 ± 3.1
128	$74,403.9 \pm 0.0$	260.6 ± 3.9	$30,674.1 \pm 0.0$	262.9 ± 5.1
256	$78,271.1 \pm 5,728.3$	223.8 ± 9.1	$58,278.9 \pm 0.0$	265.3 ± 2.3

ESMC 300M

	Full		QLoRA	
Batch Size	**Memory (MB)**	**Power (W)**	**Memory (MB)**	**Power (W)**
1	$33,306.6 \pm 42,823.0$	133.6 ± 40.8	$20,551.0 \pm 32,277.0$	118.4 ± 30.1
2	$8,542.9 \pm 0.0$	116.7 ± 0.9	$1,934.8 \pm 0.0$	107.8 ± 0.3
4	$8,547.1 \pm 0.0$	133.1 ± 0.8	$2,045.9 \pm 0.0$	125.8 ± 0.8
8	$8,505.1 \pm 0.0$	167.2 ± 7.1	$2,293.4 \pm 0.0$	158.0 ± 3.9
16	$8,563.9 \pm 0.0$	232.3 ± 24.1	$2,784.1 \pm 0.0$	243.8 ± 5.5
32	$8,983.3 \pm 0.0$	310.2 ± 13.1	$3,658.6 \pm 0.0$	317.7 ± 4.1
64	$10,721.8 \pm 0.0$	324.8 ± 2.7	$5,615.3 \pm 0.0$	330.3 ± 1.5
128	$13,496.4 \pm 0.0$	325.6 ± 0.7	$8,903.6 \pm 0.0$	329.9 ± 1.0
256	$19,802.5 \pm 0.0$	321.9 ± 1.4	$17,369.8 \pm 0.0$	331.6 ± 0.9

ESMC 600M

	Full		QLoRA	
Batch Size	**Memory (MB)**	**Power (W)**	**Memory (MB)**	**Power (W)**
1	$14,100.3 \pm 0.0$	124.9 ± 5.0	$2,184.3 \pm 0.0$	108.7 ± 4.9
2	$14,085.7 \pm 0.0$	135.2 ± 2.2	$2,249.3 \pm 0.0$	122.9 ± 1.0
4	$14,006.0 \pm 0.0$	159.3 ± 2.0	$2,402.4 \pm 0.0$	148.8 ± 2.5
8	$14,582.7 \pm 0.0$	204.6 ± 2.9	$2,662.5 \pm 0.0$	198.1 ± 3.1
16	$13,787.9 \pm 0.0$	285.1 ± 10.6	$3,287.4 \pm 0.0$	283.3 ± 15.5
32	$14,094.0 \pm 0.0$	320.7 ± 5.1	$4,447.1 \pm 0.0$	325.2 ± 1.7
64	$15,962.6 \pm 0.0$	326.7 ± 4.1	$6,982.6 \pm 0.0$	330.9 ± 1.0
128	$20,333.1 \pm 0.0$	307.5 ± 33.7	$12,078.7 \pm 0.0$	311.4 ± 34.6
256	$29,432.6 \pm 0.0$	261.8 ± 1.3	$22,216.3 \pm 0.0$	270.8 ± 1.2

ProtBERT

	Full		QLoRA	
Batch Size	**Memory (MB)**	**Power (W)**	**Memory (MB)**	**Power (W)**
1	$9,822.1 \pm 0.0$	112.0 ± 2.9	$1,829.9 \pm 0.0$	101.0 ± 0.8
2	$9,822.1 \pm 0.0$	121.6 ± 2.4	$1,888.6 \pm 0.0$	107.5 ± 1.1
4	$9,857.8 \pm 0.0$	135.2 ± 4.0	$2,014.4 \pm 0.0$	115.5 ± 5.7
8	$9,883.0 \pm 0.0$	169.3 ± 3.9	$2,217.9 \pm 0.0$	144.9 ± 1.2
16	$10,128.3 \pm 0.0$	235.9 ± 0.5	$2,654.1 \pm 0.0$	206.6 ± 5.4
32	$10,239.5 \pm 0.0$	299.0 ± 5.8	$3,522.3 \pm 0.0$	277.4 ± 9.3
64	$11,481.0 \pm 0.0$	305.2 ± 0.7	$5,244.1 \pm 0.0$	312.5 ± 1.0
128	$14,043.7 \pm 0.0$	309.1 ± 0.8	$8,740.0 \pm 0.0$	313.3 ± 2.5
256	$19,590.7 \pm 0.0$	300.8 ± 10.2	$15,723.5 \pm 0.0$	320.7 ± 4.7

ProtT5

	Full		QLoRA	
Batch Size	**Memory (MB)**	**Power (W)**	**Memory (MB)**	**Power (W)**
1	$58,910.2 \pm 0.0$	189.6 ± 14.5	$6,017.9 \pm 0.0$	139.8 ± 10.3
2	$59,772.1 \pm 0.0$	218.8 ± 18.6	$6,225.5 \pm 0.0$	172.9 ± 6.5
4	$60,160.1 \pm 0.0$	267.6 ± 19.0	$6,628.2 \pm 0.0$	224.5 ± 7.4
8	$59,847.6 \pm 0.0$	311.3 ± 4.5	$7,429.3 \pm 0.0$	279.3 ± 10.6
16	$60,009.1 \pm 0.0$	318.9 ± 1.0	$8,897.3 \pm 0.0$	308.5 ± 1.9
32	$59,702.9 \pm 0.0$	321.0 ± 1.6	$12,135.3 \pm 0.0$	316.2 ± 0.8
64	$63,112.9 \pm 0.0$	321.9 ± 0.7	$18,412.1 \pm 0.0$	315.8 ± 2.9
128	$78,745.0 \pm 0.0$	320.5 ± 2.6	$31,141.8 \pm 0.0$	315.5 ± 0.1
256	$78,768.0 \pm 18.4$	240.5 ± 25.2	$59,638.3 \pm 4,292.6$	268.2 ± 40.0

Ankh-Base

	Full		QLoRA	
Batch Size	**Memory (MB)**	**Power (W)**	**Memory (MB)**	**Power (W)**
1	$17,831.2 \pm 0.0$	114.5 ± 4.6	$2,568.1 \pm 0.0$	100.5 ± 0.3
2	$17,785.0 \pm 0.0$	125.6 ± 8.2	$2,670.9 \pm 0.0$	112.4 ± 4.5
4	$17,699.0 \pm 0.0$	153.3 ± 9.7	$2,918.3 \pm 0.0$	123.7 ± 1.3
8	$17,642.4 \pm 0.0$	245.8 ± 2.2	$9,066.3 \pm 9,842.3$	157.8 ± 2.3
16	$37,346.3 \pm 0.0$	267.8 ± 0.7	$24,178.3 \pm 0.0$	238.6 ± 2.4
32	$25,194.0 \pm 11,452.3$	273.6 ± 3.0	$6,095.5 \pm 0.0$	274.6 ± 1.1
64	$21,889.2 \pm 0.0$	276.6 ± 0.5	$9,723.6 \pm 0.0$	275.2 ± 1.6
128	$29,055.1 \pm 0.0$	273.5 ± 13.6	$17,057.3 \pm 0.0$	269.3 ± 3.2
256	$44,639.1 \pm 0.0$	264.7 ± 7.1	$31,632.5 \pm 0.0$	273.6 ± 4.1