

Ring Attention for Long-Context Protein Structure Prediction

Aaron Jarmusch^{1,2}, Tal Ben-Nun¹, Brian Van Essen¹

Lawrence Livermore National Laboratory¹, University of Delaware²



Predicting structures for very long proteins is often limited less by biology than by the memory cost of attention, forcing long targets to be cropped or skipped entirely. NEBU_HUMAN, a 8,525-residue protein, does not fit within OpenFold's current implementation. We implemented a ring-attention architecture into OpenFold, an open-source deep learning-based approach for protein folding, extending the reachable scale of structure prediction to proteins that previously could not run at all with OpenFold.

INTRODUCTION

For very long proteins, the main obstacle to structure prediction is usually memory rather than biology. **Attention cost grows quadratically with sequence length**, so long targets are routinely cropped or skipped. That can mean losing the structural layout and interactions between distant parts of the sequence that make a protein like this worth studying in the first place.

(NEBU_HUMAN) Human Nebulin at 8,525 residues exceeds what GPU memory can handle. Ring attention divides the computation into blocks passed around a ring of GPUs, so no single GPU holds the full context at once.

METHODOLOGY

Ring attention reduces the memory burden of long-context attention by splitting the sequence context across multiple GPUs instead of requiring each device to hold the full attention matrix.

1. Partition the sequence
2. Circulate context around a ring
3. Accumulate attention updates
4. Measure the tradeoff

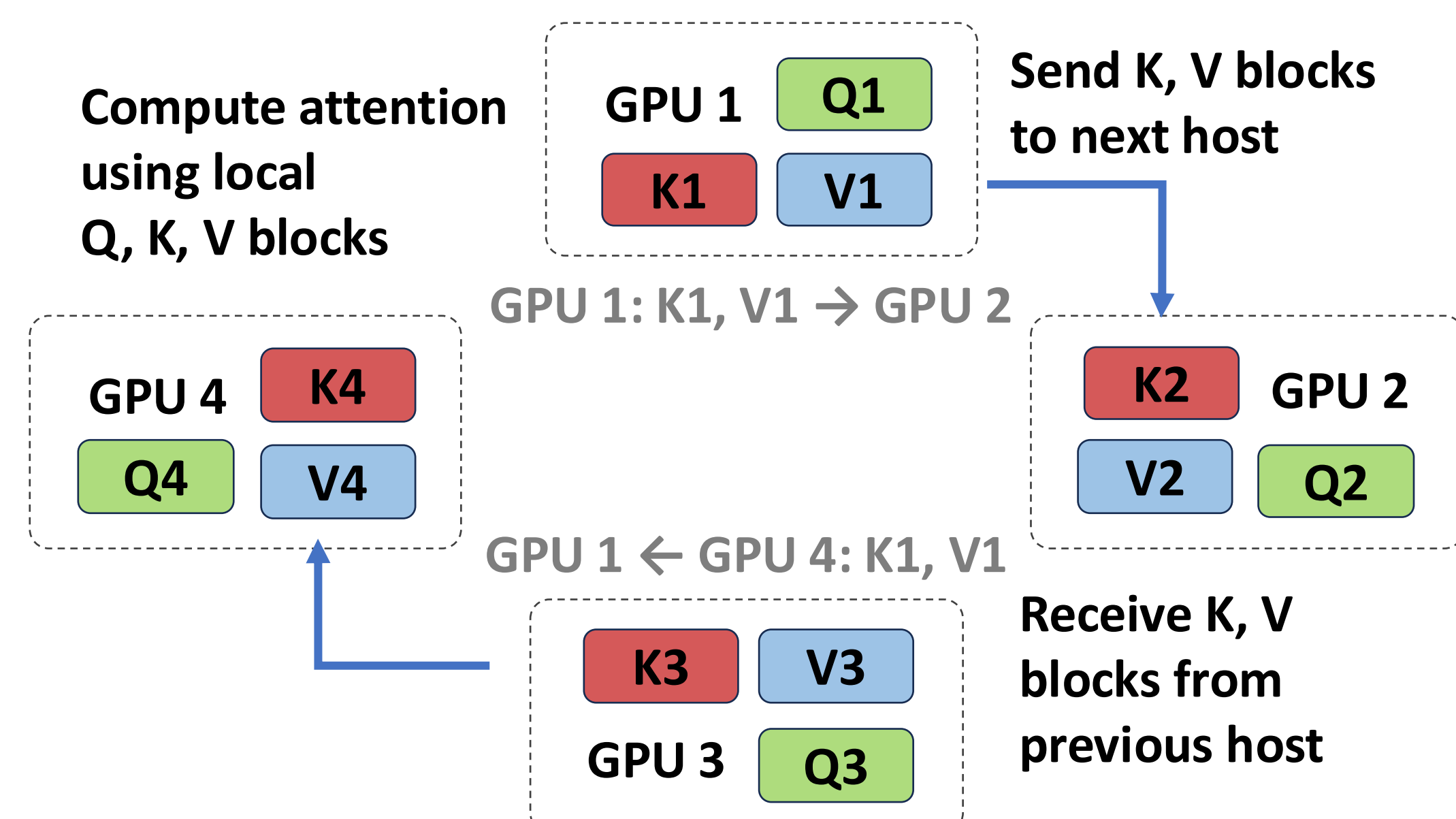


Figure 2: The K and V movement in ring attention.

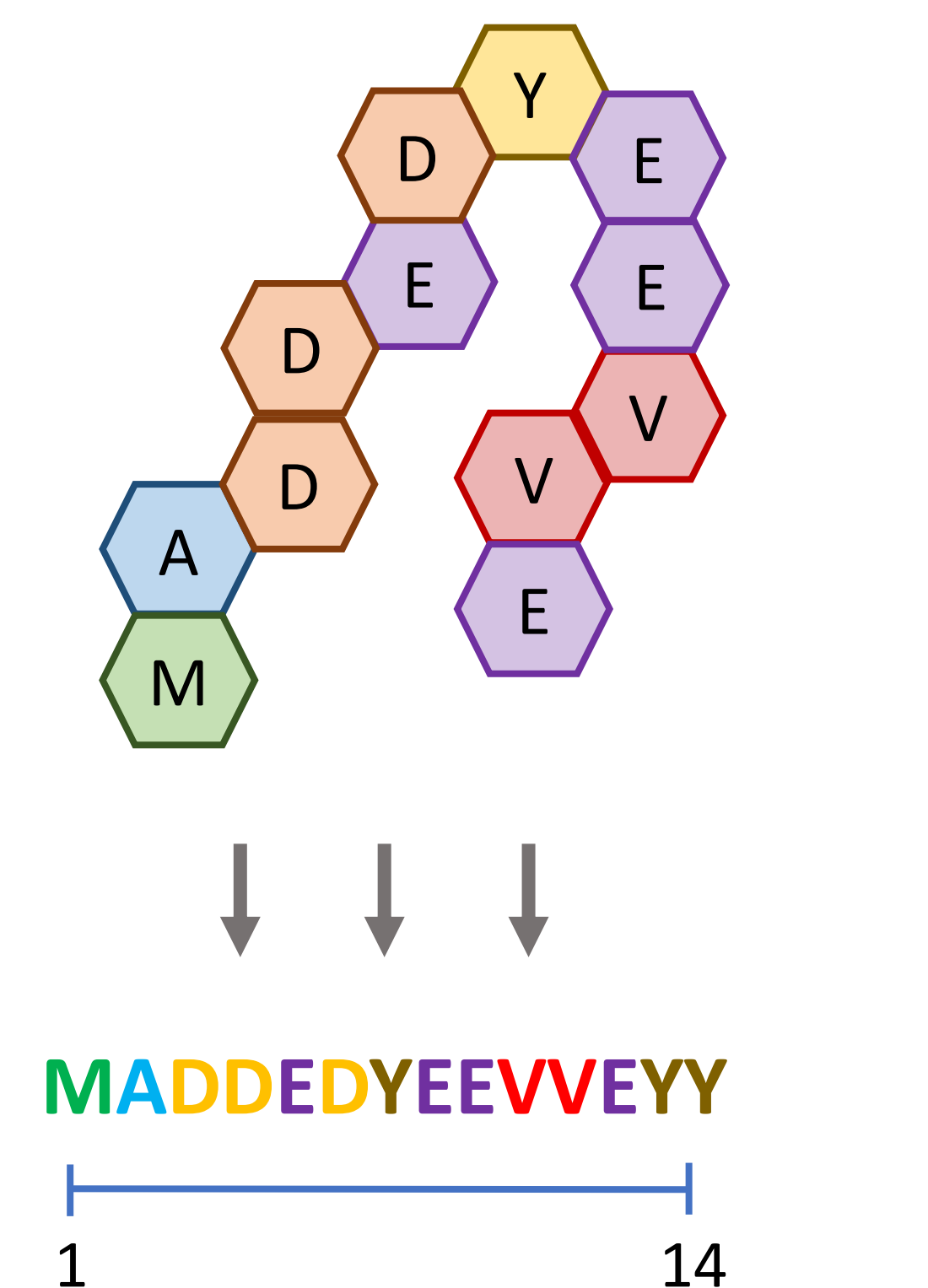


Figure 1: Length of protein sequences

CASE STUDY PROTEIN: NEBU_HUMAN

Human Nebulin is one of the largest proteins in the human body, second only to titin, it is found in the skeletal muscle, where it runs alongside the actin-containing thin filaments of the sarcomere.

It is thought to act as a "molecular ruler" that sets their length during muscle development.

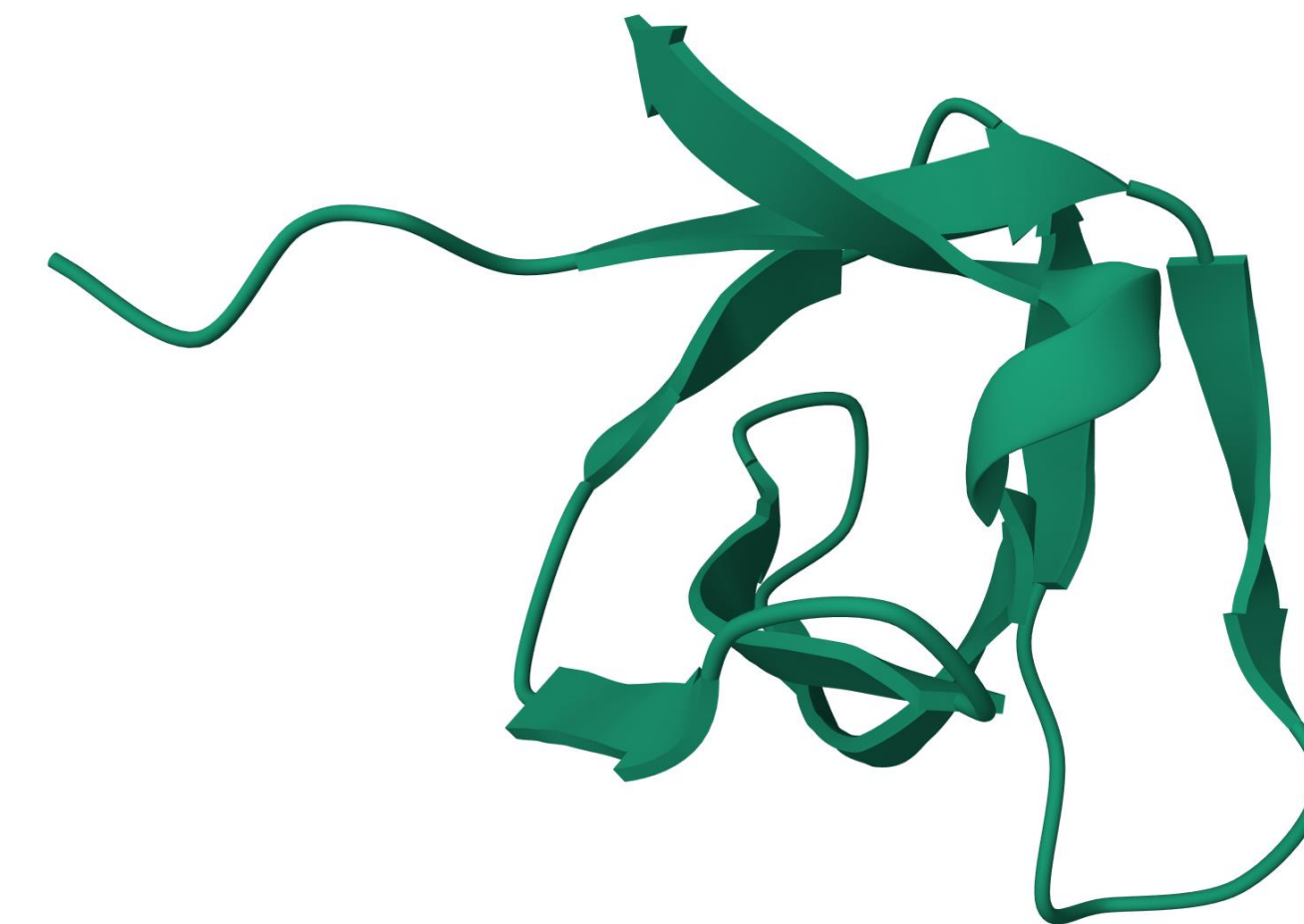


Figure 3: NEBU_HUMAN Folded Structure

EVIDENCE

Setting	Full NEBU_HMN?	Peak Memory	Status
Baseline Attention	NO	OOM/1109 GiB	Fails
Ring Attention	YES	18.8 GiB	Completed

Table 1: Feasibility Table

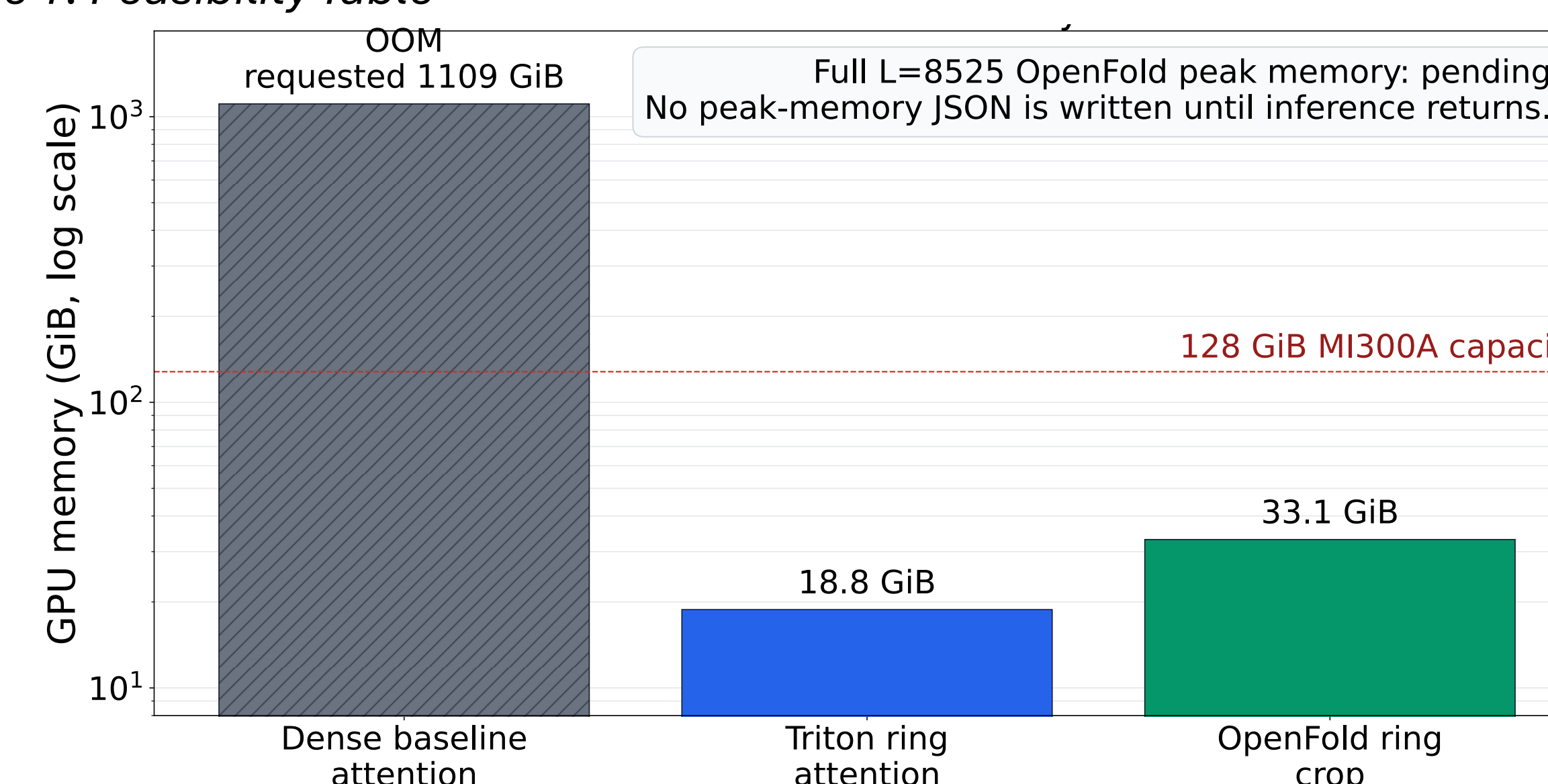


Figure 4: NEBU_HUMAN Memory Bar

- NEBU_HUMAN provides a direct test of whether ring attention can make long-protein prediction practical.
- Key Questions: how much peak memory is reduced, and whether the resulting confidence remain across the repeated regions of the protein.

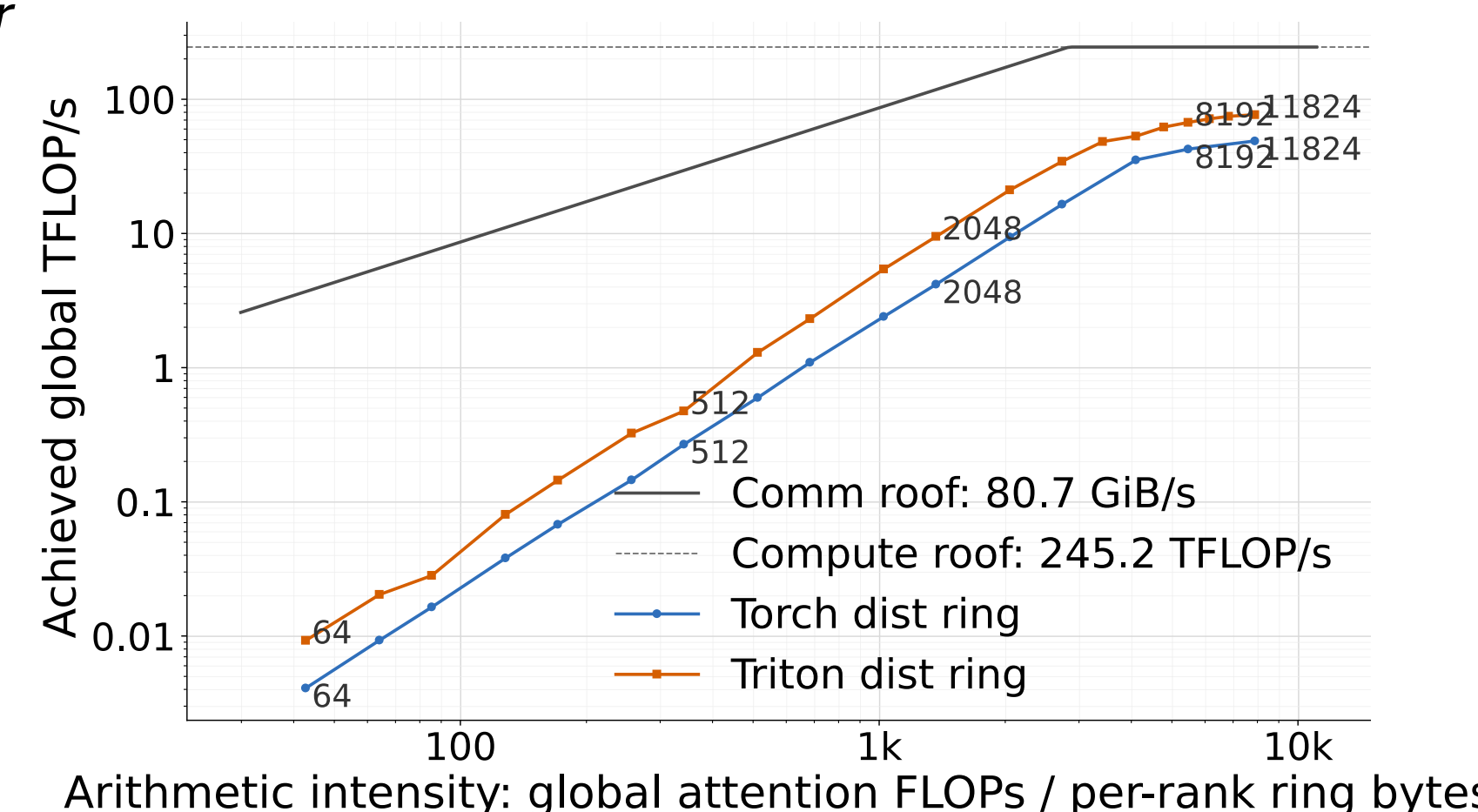


Figure 5: Triton roofline analysis of ring attention.

FULL-LENGTH PREDICTION

Ring attention enables a full-length NEBU_HUMAN prediction that is not possible with the baseline OpenFold attention setup in our current configuration.

The predicted structure shows an elongated architecture. This output is best interpreted as a feasibility result and a starting point for evaluating long-range organization, not as a finished high-resolution structure.

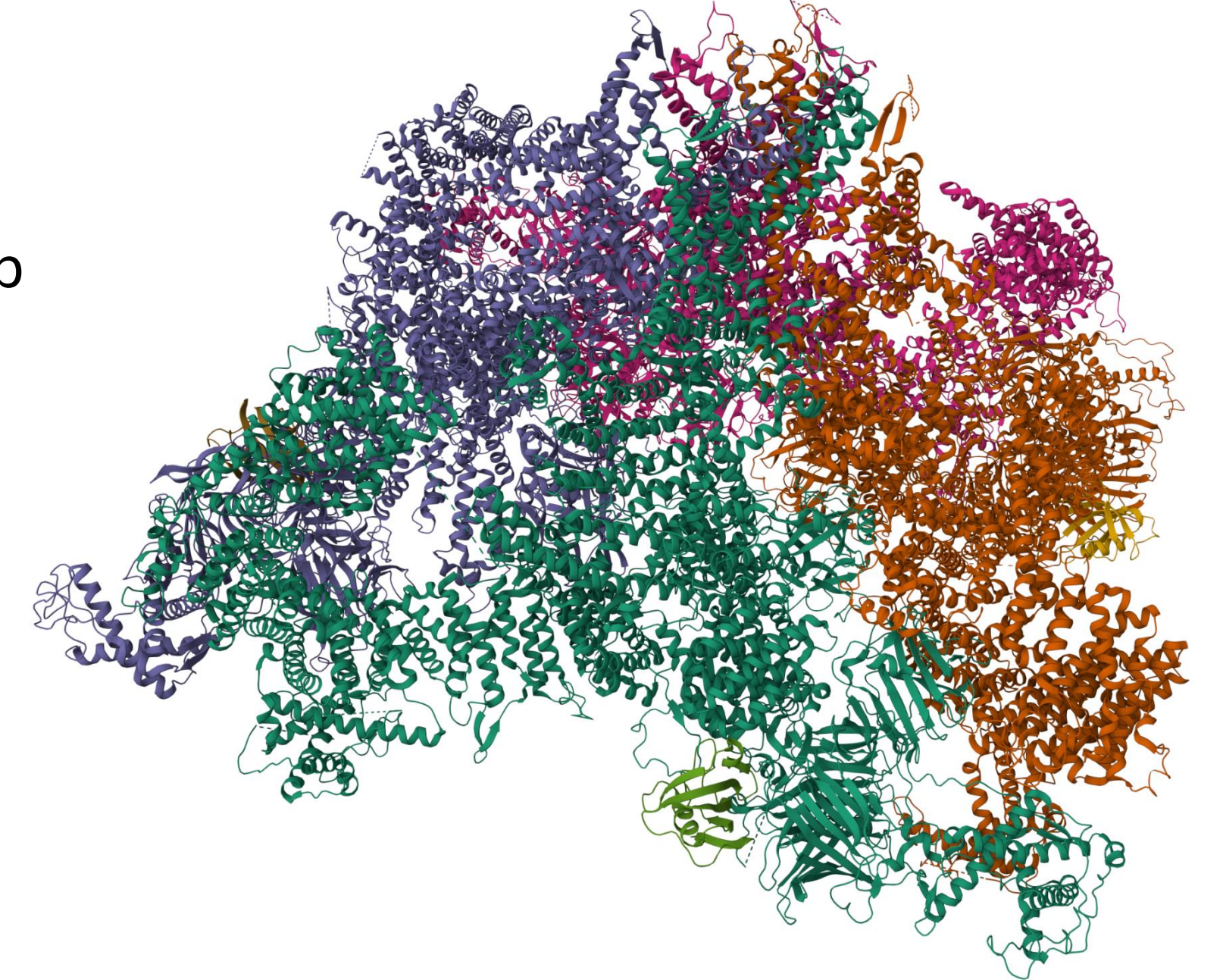


Figure 5: Protein Folded Structure not NEBU_HUMAN

CONCLUSIONS/ FUTURE WORK

Ring attention deals with the memory bottleneck that prevents full-length NEBU_HUMAN prediction, while the Triton implementation improves the efficiency needed to make that prediction practical. Together, these results shift long-protein structure prediction from an out-of-memory failure mode to a measurable tradeoff among memory, runtime, and prediction quality.

Future work will focus on:

- measuring memory and runtime scaling across GPU counts
- further optimizing the Triton ring-attention implementation
- evaluating whether improved kernel efficiency translates into faster full-length structure predictions.

REFERENCES

1. Ahdriz, G. et al. "OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization." *Nature Methods* (2024).
2. Liu, H., Zaharia, M., and Abbeel, P. "Ring Attention with Blockwise Transformers for Near-Infinite Context." *arXiv* (2023).
3. Dao, T. et al. "FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness." *NeurIPS* (2022).
4. UniProt Consortium. UniProtKB: NEBU_HUMAN, Nebulin, human, P20929.
5. Tillet, P., Kung, H. T., and Cox, D. "Triton: an intermediate language and compiler for tiled neural network computations." *MAPL* (2019).

CONTACT

jarmusch1@llnl.gov jarmusch@udel.edu