ElMerFold: Exascale Distillation Workflows for Protein Structure Prediction on El Capitan

AI+ Expo

June 4, 2025

Nikoli Dryden, Brian Van Essen

Center for Applied Scientific Computing, LLNL

Prepared by LLNL under Contract DE-AC52-07NA27344.

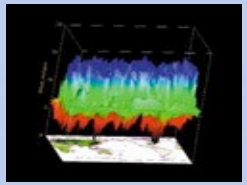
LLNL has been at the forefront of enabling HPC hardware and software since its founding

1953

1960s

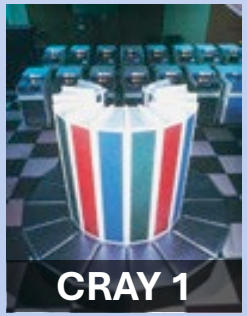
Pioneering simulations

1970s



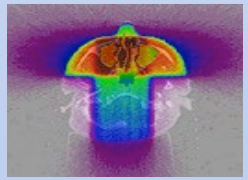
Ozone mixing models

1980s

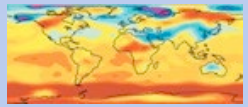


Dynamics in 3D

1990s

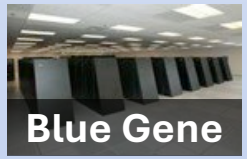


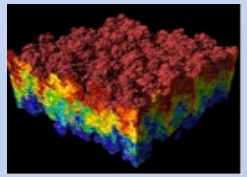
Helping the medical community plan radiation treatment



Global climate modeling

2000s

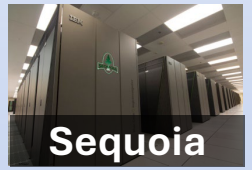


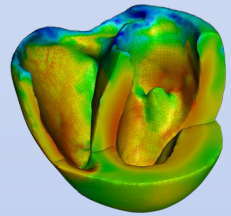


Breakthrough visualizations of mixing fluids

Unprecedented dislocation dynamics simulations

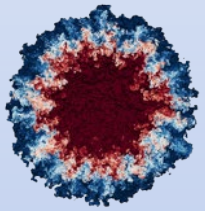
2010s





World's most detailed simulation of the human heart in action

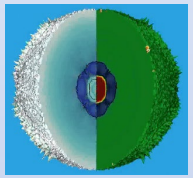
2018



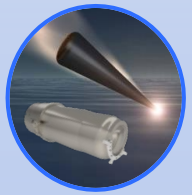
3D 98-billion-cell hydrodynamic instability simulation (ICF-inspired)

2024+

EL CAPITAN
ADVANCED SIMULATION & COMPUTING



AI guided optimization of high-yield ICF designs

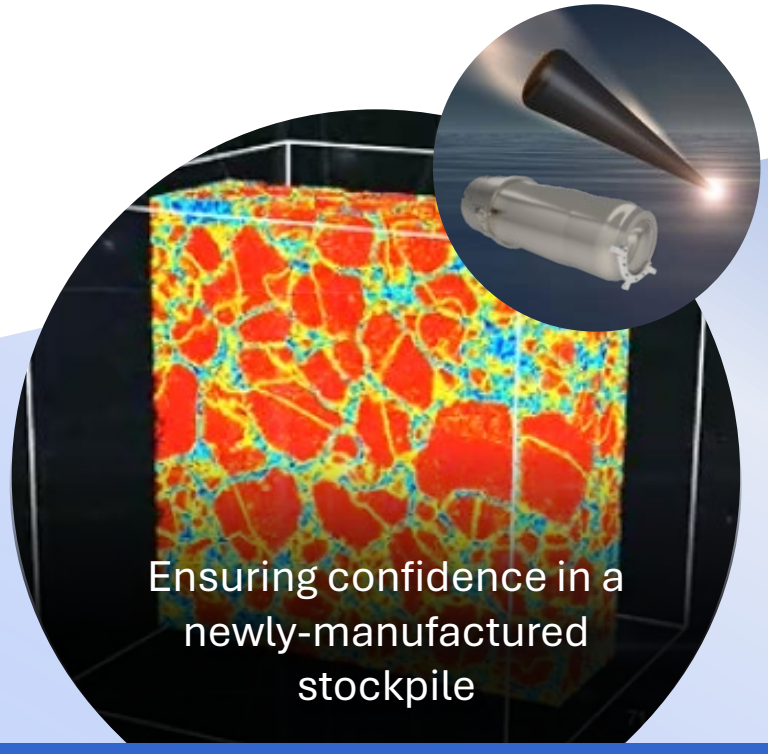


Ensembles of high-fidelity 3D weapons calculations

NNSA's Advanced Simulation & Computing (ASC) underpins the NNSA mission – past and future

AI revolution unlocked new opportunities in national security:

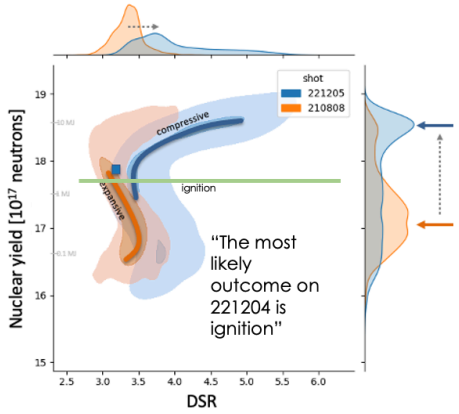
- Agent-driven design
- Cognitive simulation
- Bioresilience



Stockpile of Today	Stockpile of Tomorrow	Stockpile of the Future
W88 W80 W76 B83 W87 W78 B61	W80-4 W87-1 W93	Phase 0/1 studies

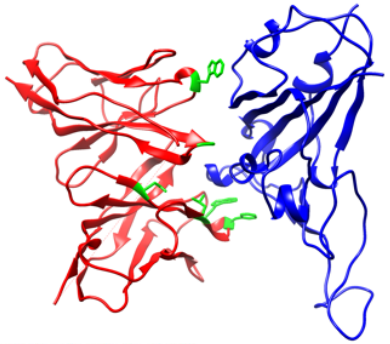
LLNL is delivering mission impact with AI expertise

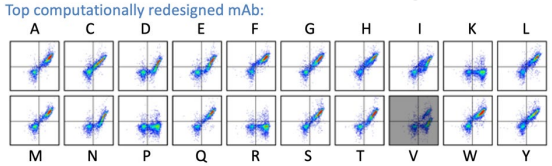
Fusion ignition



AI-driven prediction of critical systems

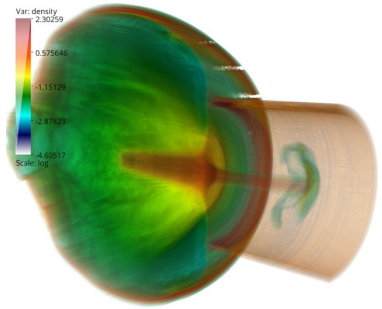
Biodesign

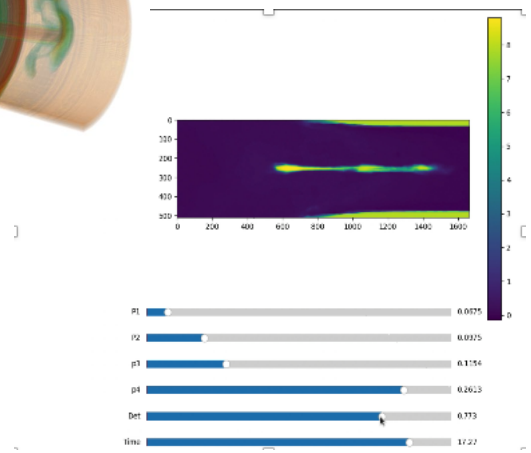




AI acceleration for biodefense and national security materials

Weapon design and certification





AI-enabled innovation and production

We must now scale this up for transformation of our critical missions

AI enables accelerated response for emerging bio threats: New approaches are needed for rapid antibody development

TRADITIONAL

Recovery

Acquire serum, screening

Antibody engineering
Repeat if fail pre-clinical

- Requires convalescent serum
- Manufacturing challenges
- Produces narrow protection

4–6 months

FUTURE

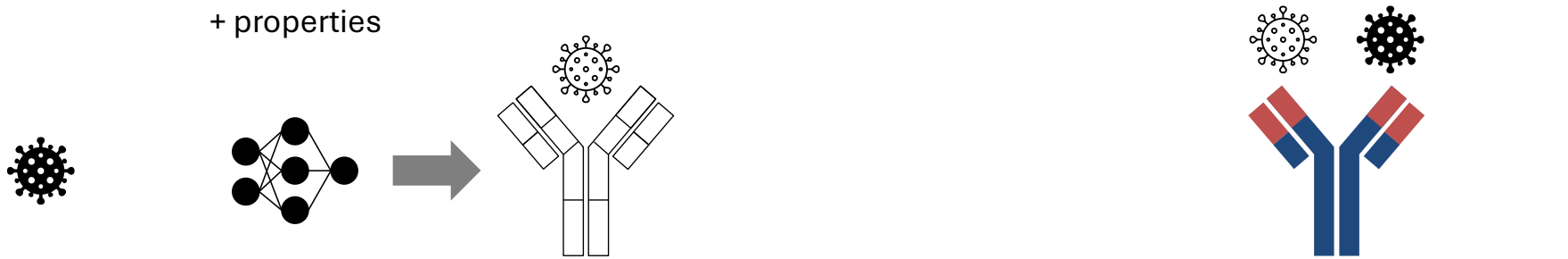
Emerging

Generative model to optimize design + properties

Prototype

In silico + experimental optimization loop

Cross-reactive, manufacturable antibody



- Machine learning models predict antibody design and properties
- High-quality protein folding model is critical
- Explore design space in silico, optimizing for quality
- Manufacture and test most promising designs

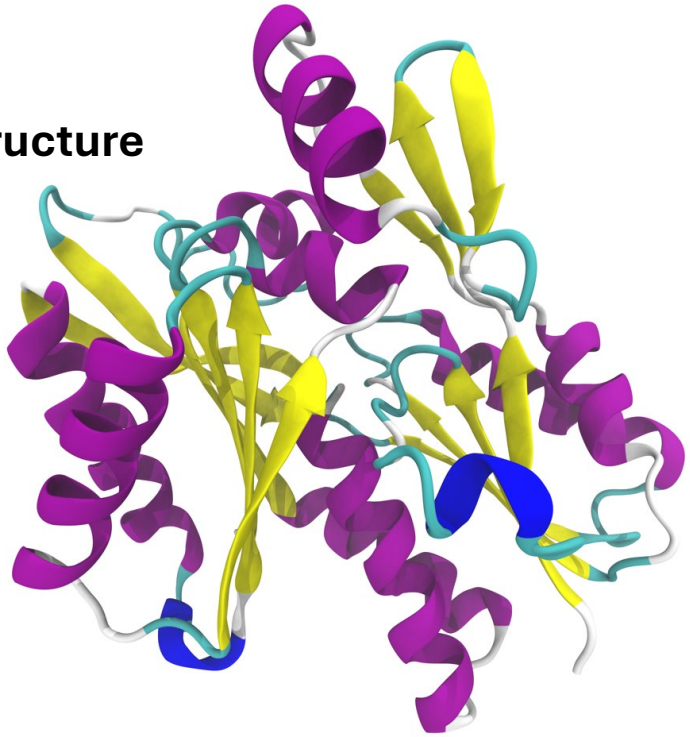
Target: ~4 weeks (or less)

Understanding how proteins fold is the basis for many biological applications

Sequence

MSYKIIGDSCLDLTEELKKDSRFATVSLTLQVDDTMVI
DDDTFDQKAFLDLVKASENCPKSACSSPDAFKQAM
ECDEDDVYIITLSSHLSGSYNSAVIGKELYEEEHGAD
HKNILVIDSESASAGELNLALGICEMYEEGLDFQAIS
EKIMKKRDDENIYFVLDTLDFLRKNGRLTGLQAFFAT
ALNIKPIMGADKGVIIKLDQARGINKAFAKMCEFAVK
EAGESEKKRVIIAHCNCPERAAQVKQELEKRASFREI
LITETAGVATMYAGDGGVILSIEG

Structure



MGYP000021002699, OpenFold 2 prediction, mean pLDDT 96.3

Computational prediction of protein structure enables many applications

Deep learning has revolutionized protein folding

Jumper et al., 2021:

Article

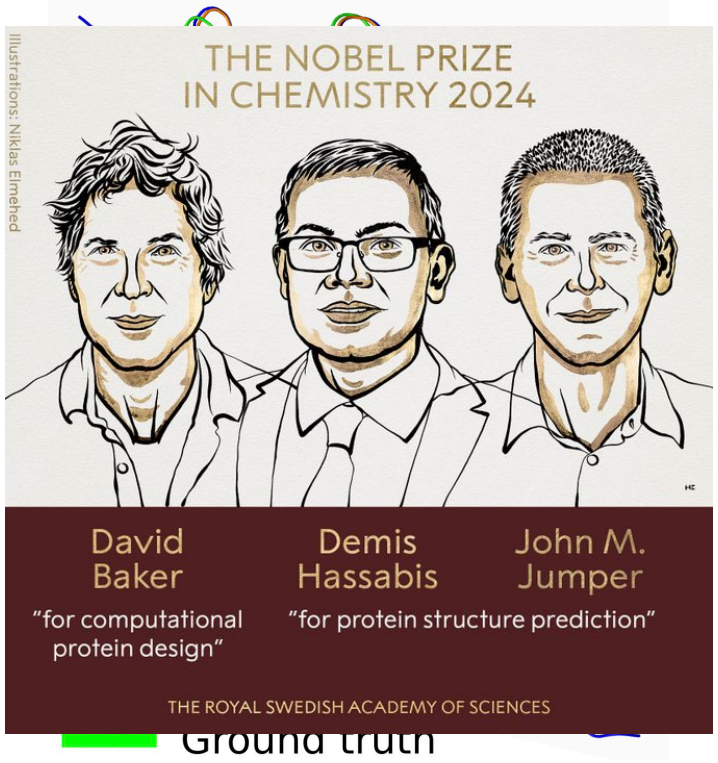
Highly accurate protein structure prediction with AlphaFold

"End-to-end folding by AlphaFold 2 largely solved the single domain protein structure prediction problem" (Pearce & Zhang, *J Biol Chem*, 2021)

Abramson et al., 2024:

Article

Accurate structure prediction of biomolecular interactions with AlphaFold 3

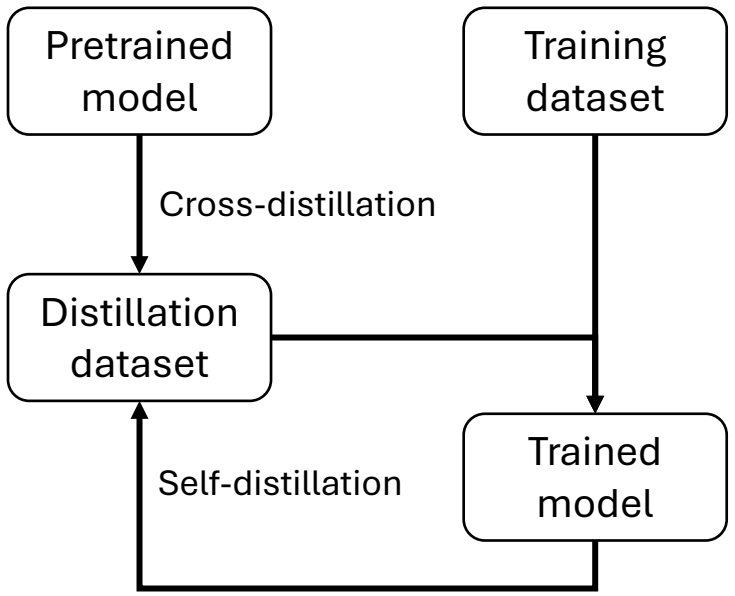


Experimental datasets are not sufficient — Distillation datasets are critical for protein folding models

Distillation: Using a pretrained model to produce **high-quality** data for further training

- Self-distillation & cross-distillation
- Calibrated scoring of predictions is critical
- Anchored by experimental results

Major component of AlphaFold 3



Shen et al., 2024

Jumper et al., 2021

For national security, it is critical for the USG to be able to evolve state-of-the-art models: Collaboration to train OpenFold 3

- AlphaFold 3 is the state-of-the-art
  - Training code and some datasets unavailable
  - Very restrictive license (distillation prohibited!)
- Existing reproductions (e.g., Boltz) use substantially less compute and data
- Open models spur innovation and enable new insights
- Retraining is critical for improving models and addressing future needs

openfold

Democratizing AI for Biology

OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization

Gustaf Ahdritz, Nazim Bouatta, Christina Floristean, Sachin Kadyan, Qinghui Xia, William Gerecke, Timothy J. O'Donnell, Daniel Berenberg, Ian Fisk, Niccolò Zanichelli, Bo Zhang, Arkadiusz Nowaczynski, Bei Wang, Marta M. Stepniewska-Dziubinska, Shang Zhang, Adegoke Ojewole, Murat Efe Guney, Stella Biderman, Andrew M. Watkins, Stephen Ra, Pablo Ribalta Lorenzo, Lucas Nivon, Brian Weitzner, Yih-En Andrew Ban, ... Mohammed AlQuraishi

+ Show authors

Nature Methods **21**, 1514–1524 (2024) | Cite this article

Distillation requires more compute than training

	Dataset	Description	Samples
Experimental	Monomer protein structures	Protein Data Bank structures	~200k
Distillation datasets	RNA distillation	RNA predictions from Rfam	~65k
	Disordered PDB distillation	OF2-Multimer predictions of PDB proteins	~25k
	Transcription factor distillation	DNA+protein predictions from JASPAR	~16k
	Monomer protein distillation	OF2 predictions of MGnify sequences	~41M

- Estimated OpenFold 3 training time on all of El Capitan:
  3 days (including hyperparameter tuning)
- Estimated time to produce the monomer protein distillation data on all of El Capitan:
  >1 week
- **Distillation data generation dominates training time**

El Capitan: Flagship NNSA Tri-Lab capability for exascale computational science

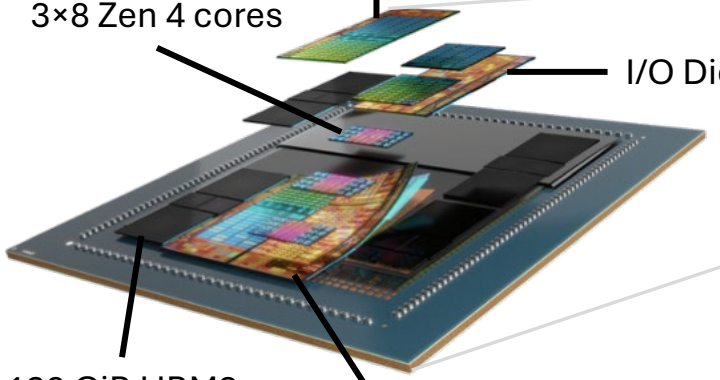
- #1 on November 2024, Top500 list: 1.74 Eflop/s on High Performance Linpack (HPL)
- 11,136 nodes with 4 AMD MI300A Accelerated Processing Units (APUs) each (44,544 total)
- HPE Slingshot interconnect (3-level Dragonfly), 4 Cassini NICs per node (100 GiB/s)
- Capacity tier storage: Merced, 401 PiB HDD-backed Lustre filesystem
- Performance tier storage: 696 HPE Rabbit modules, 21 PiB configurable NVMe storage
- Tuolumne: Open science “sister” of El Capitan (1,152 nodes, 4,608 APUs total)

El Capitan system architecture highlights

Accelerator Complex Die (XCD)
6×38 AMD CDNA3 CUs

CPU Complex Die (CCD)
3×8 Zen 4 cores

I/O Die

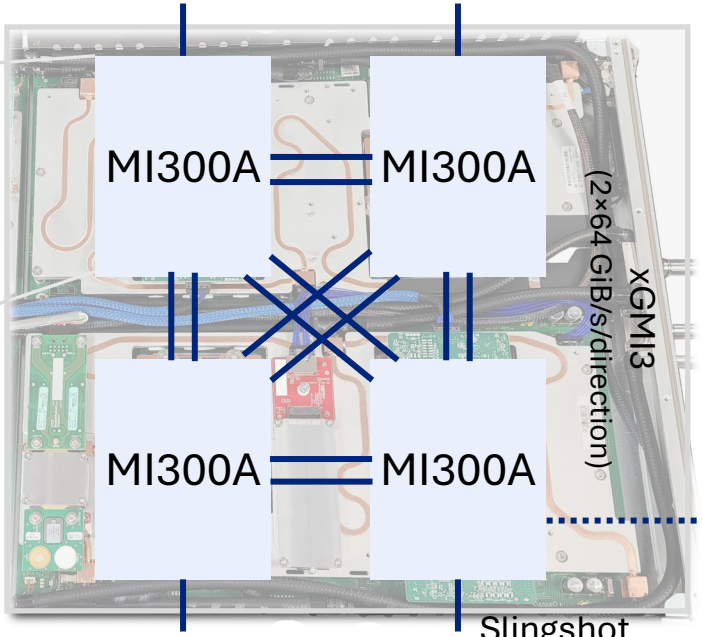


128 GiB HBM3
(8×8H)

Infinity Fabric

Achieved FP32 peak (SGEMM): 82.6 Tflop/s
Achieved mem bandwidth (STREAM): 3.2 TiB/s

Compute node



Slingshot
(25 GiB/s/direction)

Merced Lustre filesystem
401 PiB (on Slingshot fabric)

Rabbit (near-node storage)

16×2 TiB NVMe SR-IOV
(plus 2× spares)

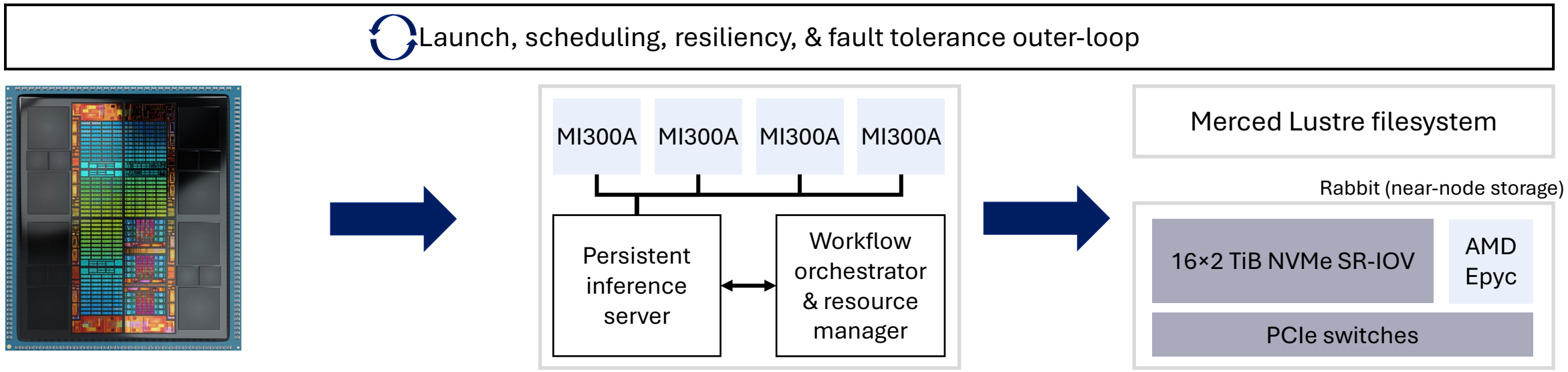
AMD
Epyc

PCIe switches
(2×96P)

PCIe-G4 x4 links to 16 compute nodes
(4 GiB/s/direction)

ElMerFold: Scaling distillation data generation to El Capitan

High-performance framework for large-scale inference on El Capitan + OpenFold-specific optimizations



MI300A optimizations

- LBANNv2: Memory optimizations
- Efficient kernel implementations

Triton DaCe

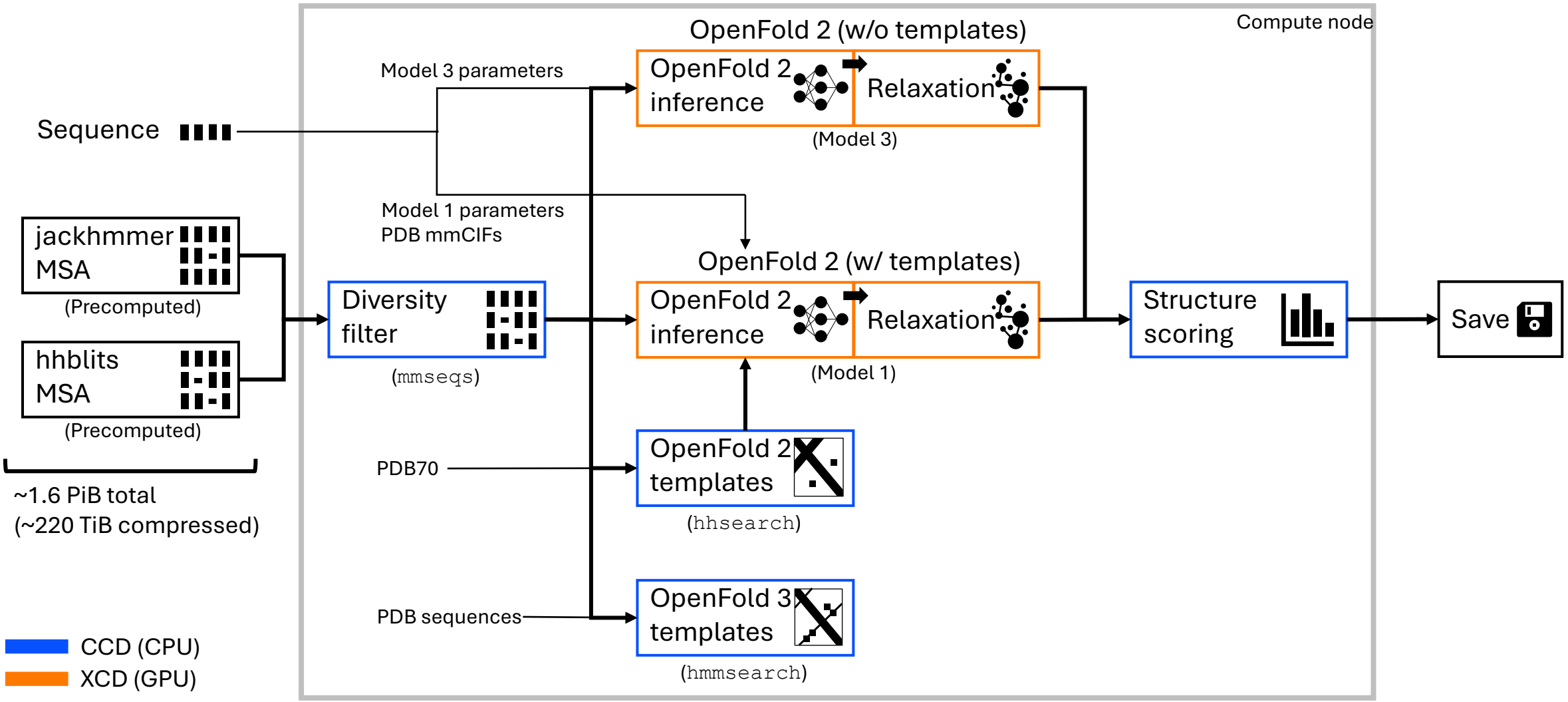
On-node optimizations

- Workflow orchestration
- Persistent inference servers

I/O optimizations

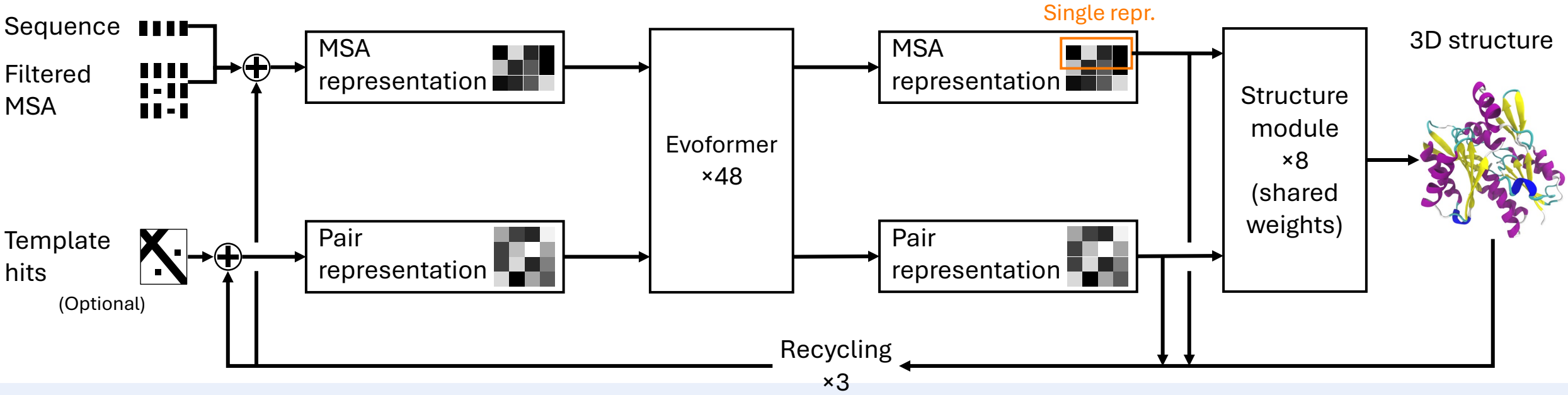
- I/O offload to Rabbits
- Streaming data (de)compression
- Startup optimizations SPINDLE

OpenFold 3 monomer distillation workflow



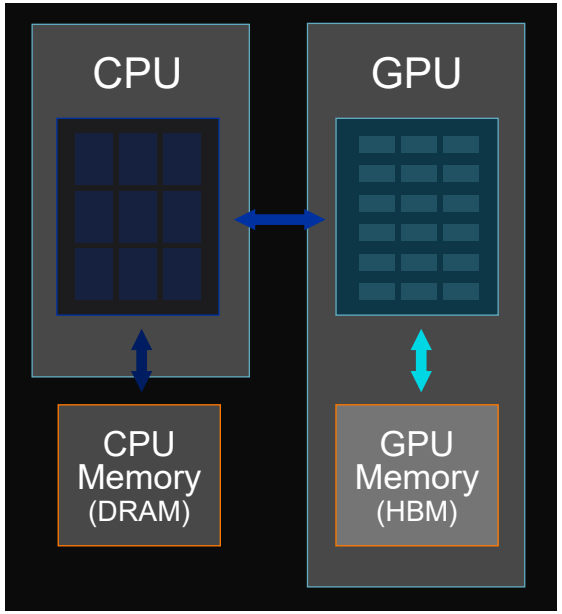
OpenFold 2 is a complex model

- Transformer-based, but with multiple non-standard attention mechanisms (EvoFormer)
- Typical inference time is seconds to minutes per sequence
- Runtime is cubic in sequence length (approx. $347{,}904S^3 + 764{,}022{,}844S^2 + 300{,}965{,}502{,}304S$)

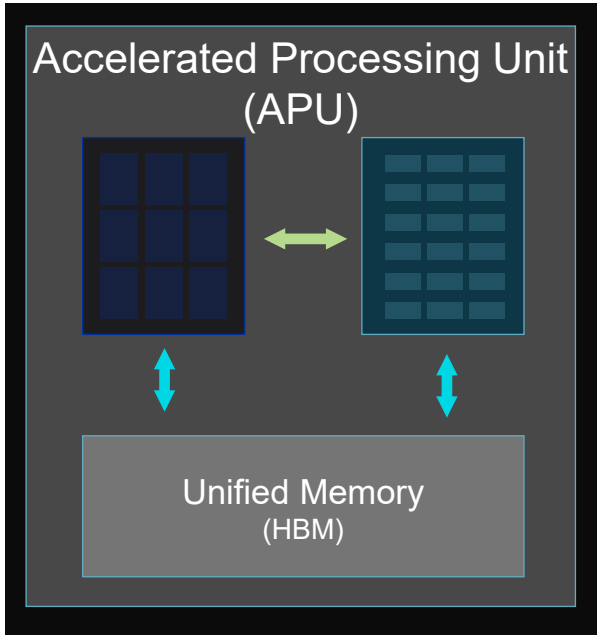


LBANNv2: Improving performance on AMD MI300A

AMD MI250X (e.g., Frontier)

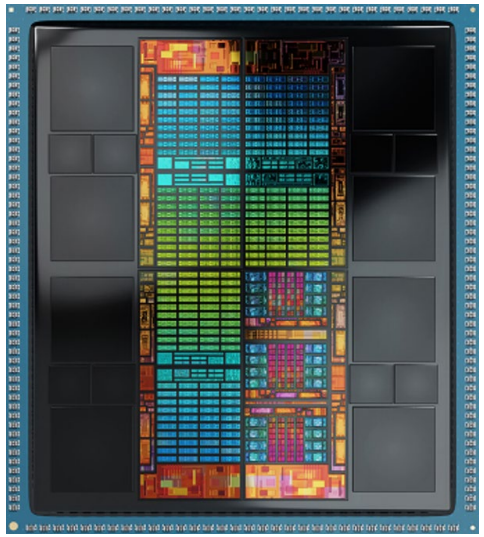


AMD MI300A



LBANNv2: Improving performance on AMD MI300A

- Elide copies between CPU and GPU devices on unified memory systems
  - Reduce memory overhead and eliminate copies!
- Requires changes to common PyTorch programming model assumptions (`tensor.to()`)



Unified HBM between CCD & XCD
No DRAM!

OpenFold 2 data pipeline (`data_pipeline.py`)**:**

```
import lbannv2
# ...
with lbannv2.MigratableMemory():
    t_result = t_f.get_templates(
        query_sequence=input_sequence,
        hits=hits_cat,
    )
    template_features = t_result.features
# ... Later:
out = lbannv2.migrate(feats, device)
```

← Initialize memory allocators & hooks

All tensors are automatically migratable between CCD and XCD devices

← Like `tensor.to()` but no copy if possible

https://github.com/LBANN

Higher-level approaches for efficient AI algorithms: Optimizing EvoFormer in OpenFold

- OpenFold 2 is well-optimized (for A100s) … but leaves performance on the table on new systems
  - Reimplementing kernels in HIP / CUDA is unsustainable and nonportable
- **Triton** provides a portable, high-level (Python) approach for writing efficient GPU kernels
  - FlashAttention-style kernels adapted to EvoFormer
- **DaCe** allows lower-level control of memory operations to improve performance

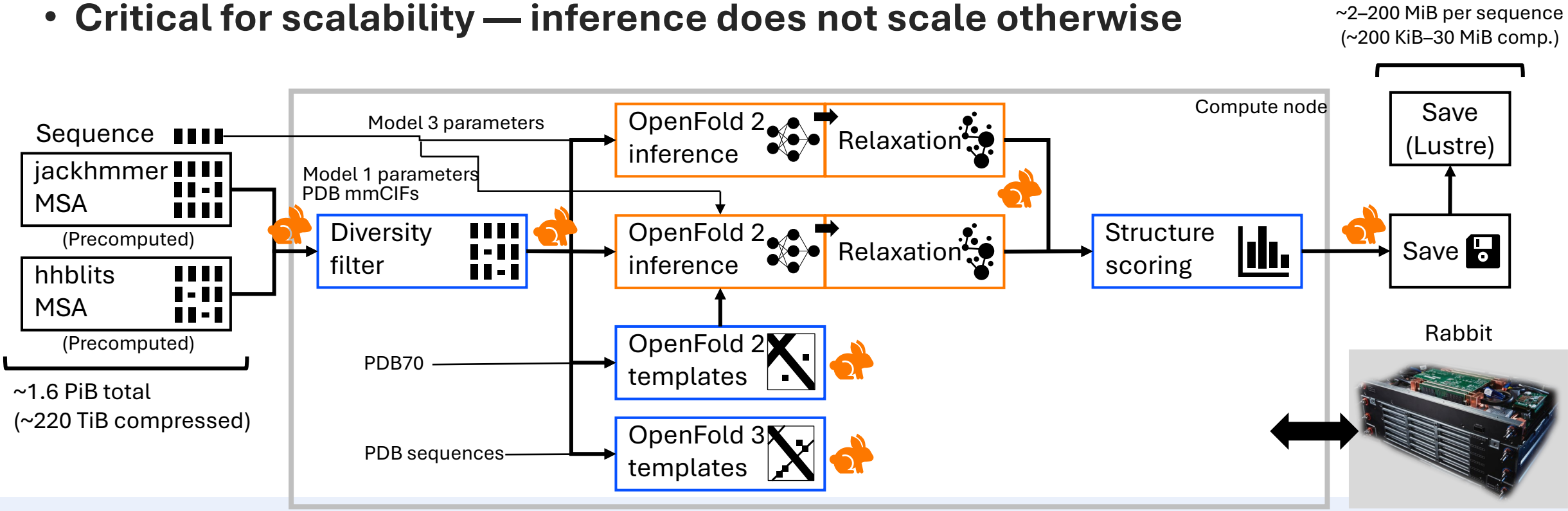
Pointwise attention kernel (#residues = 256)

	Time	Mem. read BW	Mem. write BW
Original	2.30 ms	—	—
Triton	1.12 ms	260.79 GB/s	14.49 GB/s
DaCe	167.76 µs	3,630.00 GB/s	178.84 GB/s

PyTorch Triton DaCe

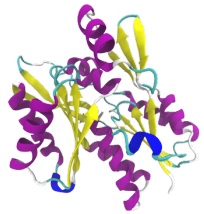
Rabbits enable efficient I/O at scale and system flexibility

- Configurable Rabbit storage allows jobs to tailor the I/O subsystem to their needs
- Every node requires ~380 GiB of static data plus storage for intermediates
- Offload intermediate data to near-node Rabbit storage
- **Critical for scalability — inference does not scale otherwise**



ElMerFold scales distillation to El Capitan

~2400 structures/s



17.2×

- Scales to 10,800 nodes (43,200 APUs) of El Capitan with **72.9%** parallel efficiency
- Delivers sustained performance of **~2400 structures/second** & **8.6M atoms/s** (including I/O and all other overheads)
- **17.2×** improvements on the OpenFold 3 monomer protein distillation dataset

Distillation datasets and large-scale inference is the ~~future~~ present

- Distillation enables training better models in data-scarce regimes
- Large-scale inference improves data quality
- Widely used for producing synthetic data already!

Abramson et al., 2024

Brown et al., 2024

DOE-led convergence of AI, high-performance computing, and automated biotech is transforming biology for national security

DOE partnership with DoD for AI Acceleration of Molecular Design for Biodefense

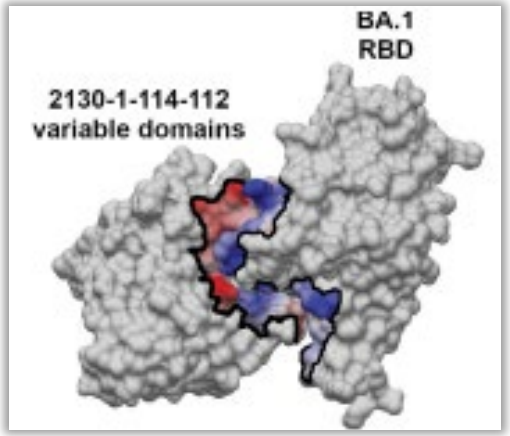
Lawrence Livermore National Laboratory

Los Alamos NATIONAL LABORATORY

AstraZeneca

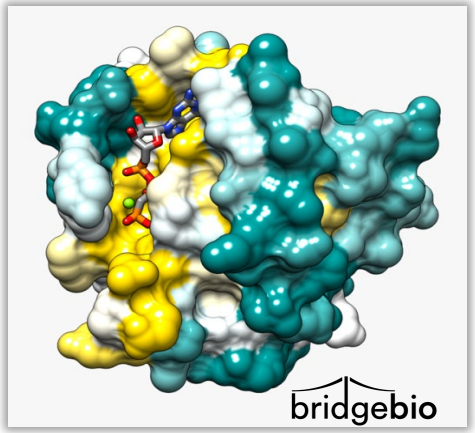
AMD

Microsoft

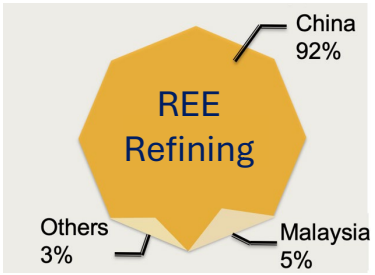


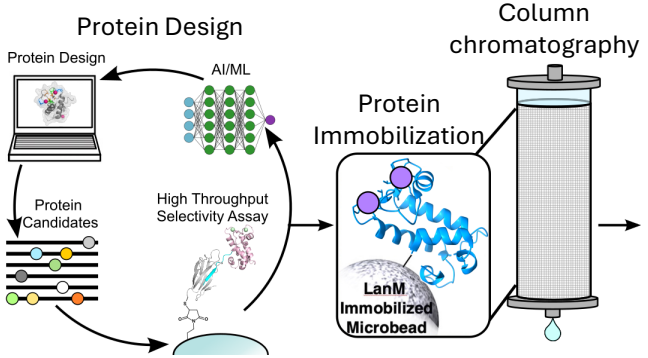
Accelerating antibody design from months to <2 weeks to ensure warfighter readiness

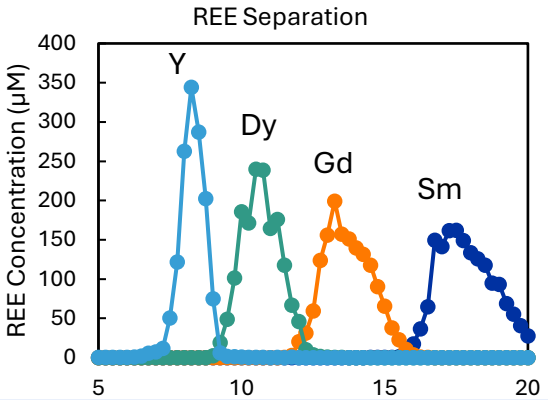
Connecting DOE capabilities to industry partners amplifies innovation → 3 new medicines in clinical trials



Harnessing AI + Biology for Efficient Rare Earth Element Separation







Alta

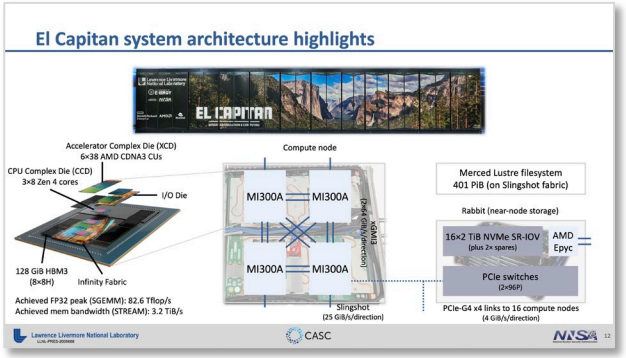
AMERICAN RARE EARTHS

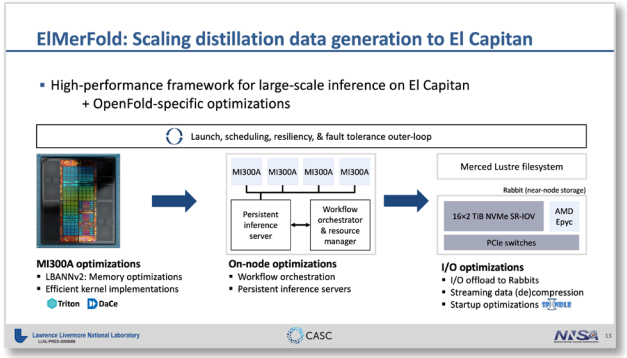
RioTinto

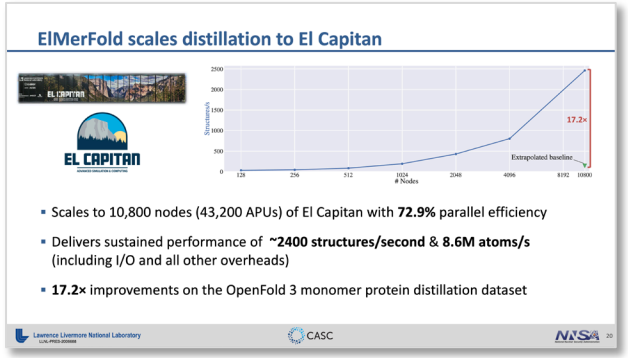
The future of scientific AI workloads will be driven by inference

Enabling large-scale inference on leadership-class HPC unlocks:

- Uncertainty quantification of AI models
- Test-time compute and reasoning agents
- Robust predictions for generative models
- Distillation-driven data sets and future model training







Calibrated model confidence and operator trust is critical to enabling adoption for high-consequence missions

The Elmer Fold Team

Nikoli Dryden

Tal Ben-Nun

Pier Fiedorowicz

Tom Benson

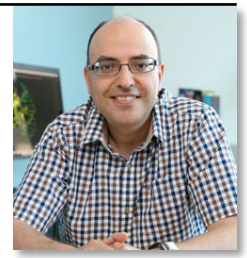
Bronis R. de Supinski

Brian C. Van Essen

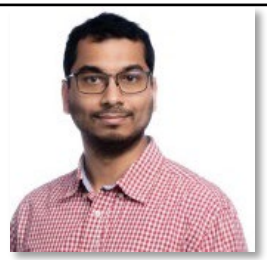
Vinay S. Swamy

Colin Kalicki

Mohammed AlQuraishi



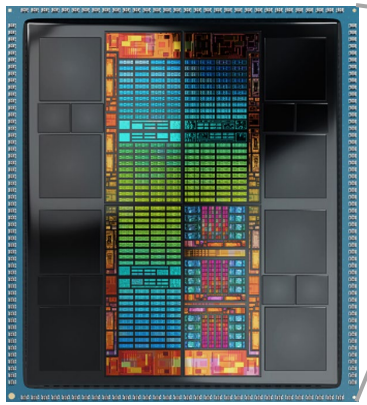
Vinayak Gokhale



J. Austin Ellis

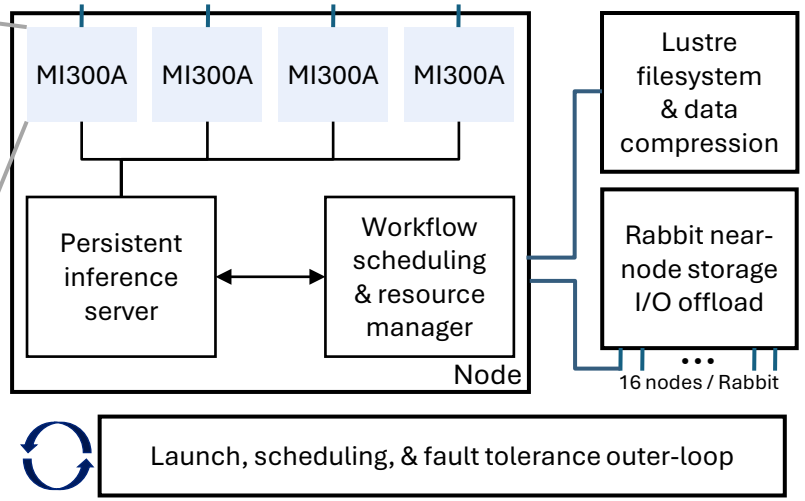
Nicholas Malaya

ElMerFold accelerates distillation dataset generation on El Capitan for NNSA missions

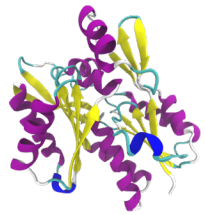


LBANNv2

Triton DaCe



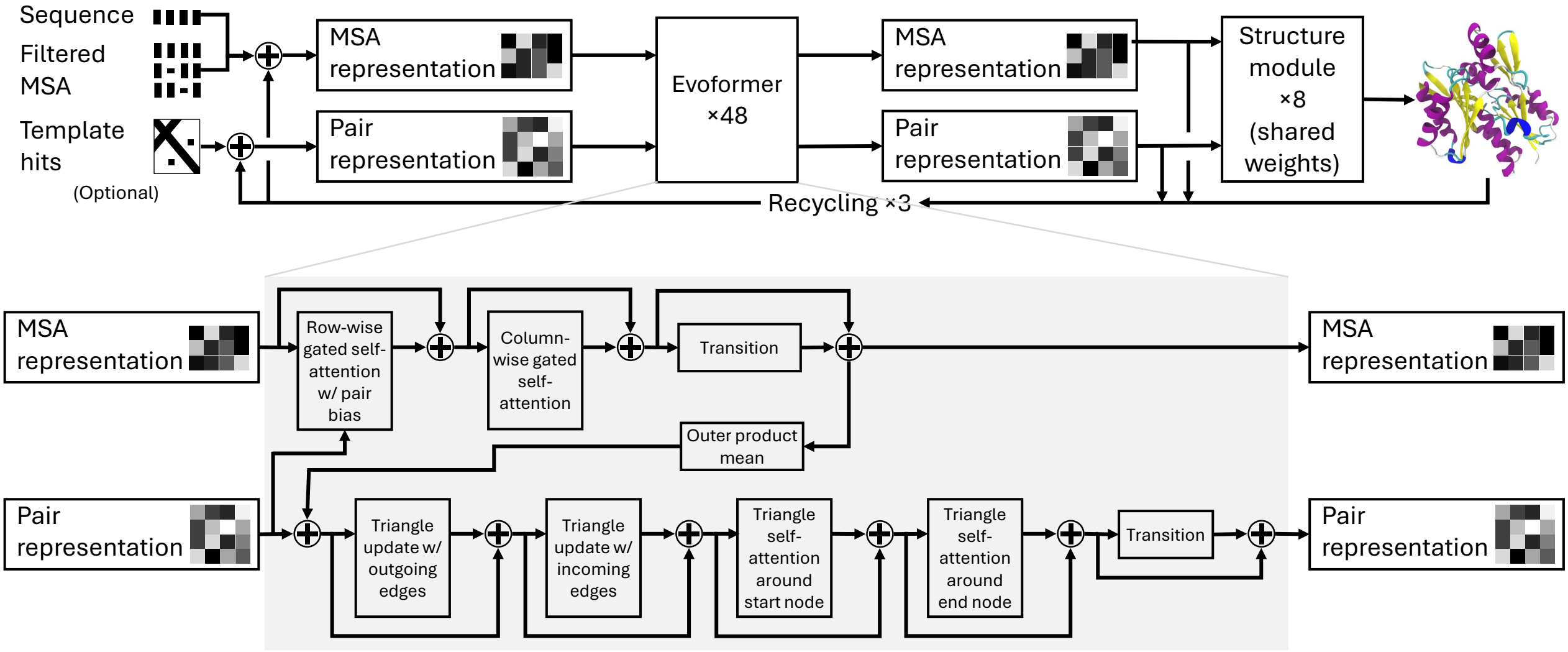
~2400 structures/s



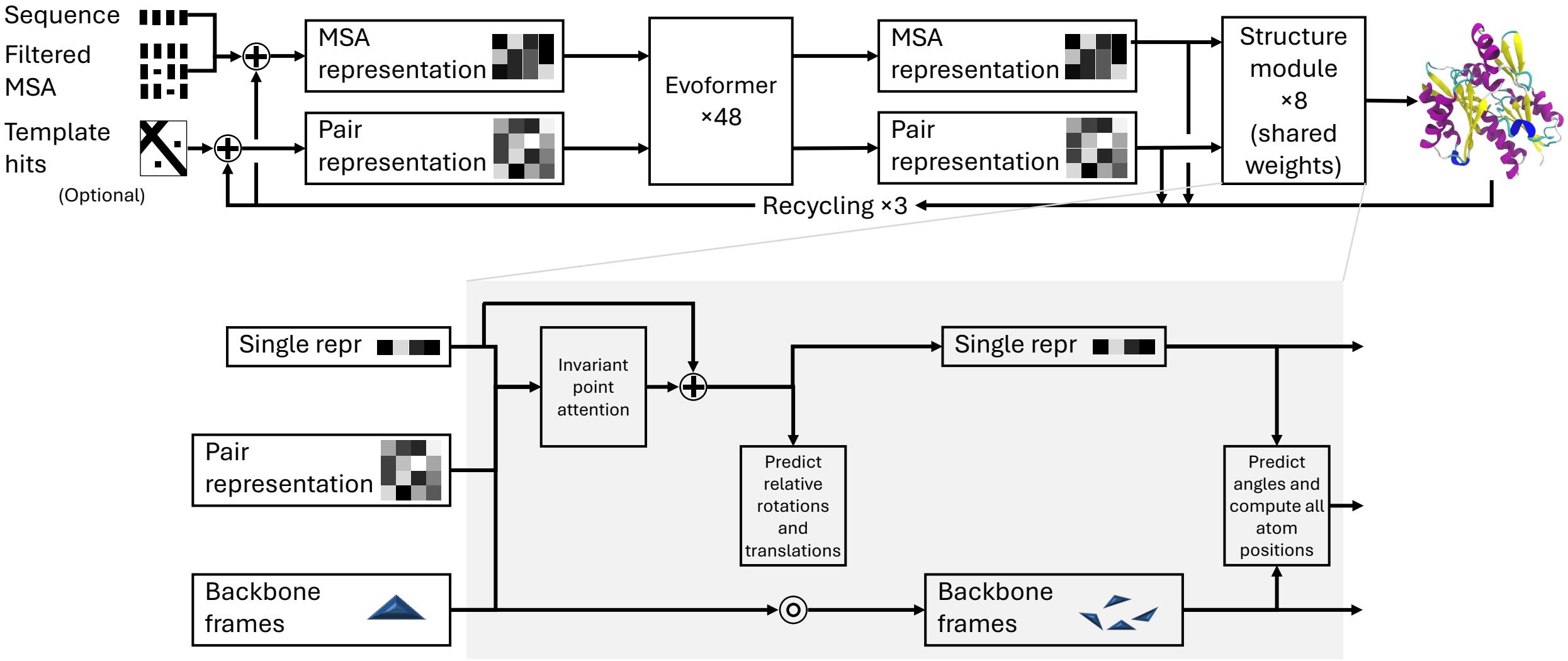
17.2×

- MI300A innovations enable leadership-class AI workloads on El Capitan
- **17.2×** improvements on the OpenFold 3 monomer protein distillation dataset at full El Capitan scale
  - Workflow scaled to 10,800 nodes (43,200 MI300A APUs)
- Unparalleled capability for generating distillation datasets enables key advances in NNSA mission needs and open science applications

OpenFold 2 is a complex model



OpenFold 2 is a complex model



On-node optimizations: Persistent inference servers and workflow orchestration

- Inference servers avoid repeated initialization & loading overhead
- Online memory management enables fine-grained oversubscription
- On-node workflow orchestration ensures efficient resource utilization