DATA SCIENCE INSTITUTE

Scaling Deep Learning for Cancer Drug Discovery on HPC Systems

Sam Ade Jacobs, Nikoli Dryden, Tim Moon, Brian Van Essen, Stewart He, Jonathan Allen

Lawrence Livermore National Laboratory

University of California National Laboratories

MOTIVATION

- Along with sources like PubChem, Genomics of Drug Sensitivity in Cancer (GDSC), and Genomics Data Common, the recently released NCI-ALMANAC database provides access to millions more cancer data samples than were previously available
- This exponential growth in size of cancer dataset makes parallel processing for large-scale deep learning an attractive tool for cancer research
- We present preliminary scaling studies of a deep learning algorithm for predicting tumor cell line response to drug pairs

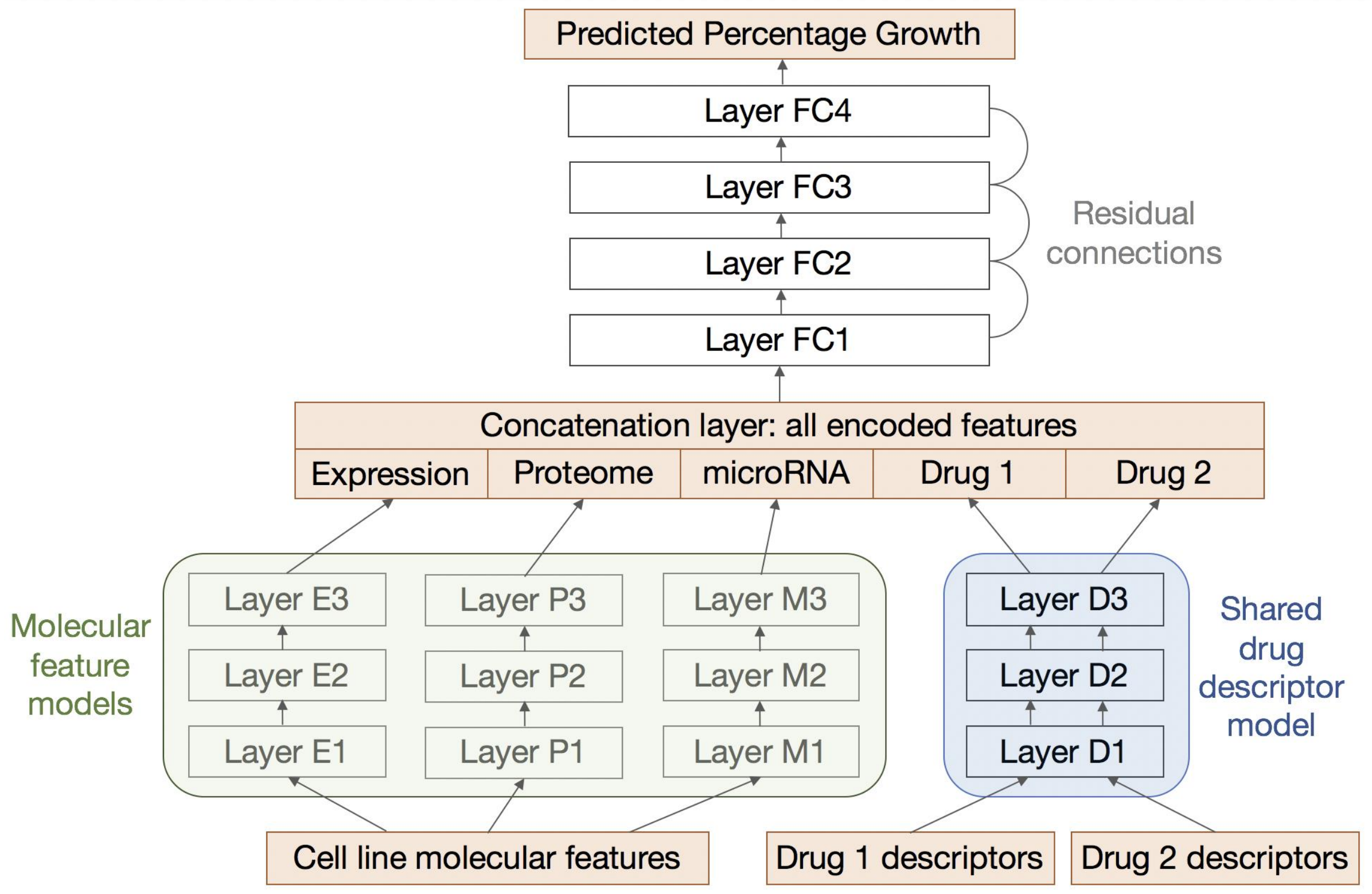
NEURAL NETWORK MODEL FOR PREDICTING TUMOR CELL LINE RESPONSE TO DRUG PAIRS

Regression Problem

- Given drug features D; descriptors, fingerprints, structure, dose etc
- And a combination of genomic cell feature τ; gene expression levels, protein abundance, microRNA expression, DNA etc
- Predictive (neural network) models quantify the response $R = f(\tau , D)$ of a given cell line τ to drug(s) D

Network Architecture

- Input is a combination of genomics features and drug descriptors
- Hidden (latent space) layers are multi-level multi-headed fully connected layers with lower level two paths consisting of drug and gene feature encoding submodels. These paths are later concatenated to feed the upper level “growth prediction” submodel
- Output is regression value typically measured with IC50 regression, G150, percentage tumor growth, or Z-score

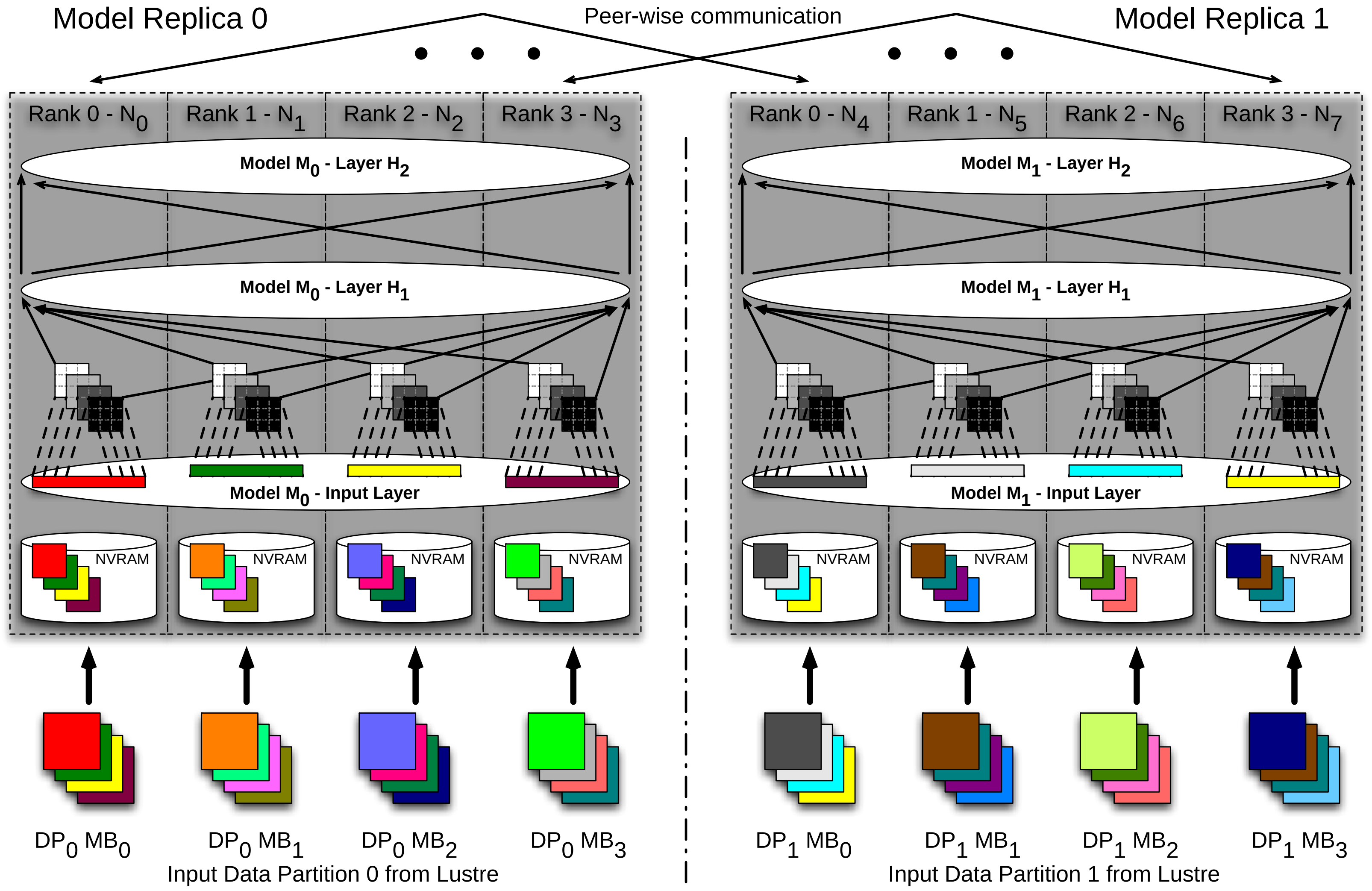


Combo model for predicting tumor cell line response to drug pairs [1]

PARALLELIZATION IN LBANN

Livermore Big Artificial Neural Network (LBANN) [2]

- HPC-centric deep learning toolkit
- Optimized distributed memory algorithm
- Optimized asynchronous all-reduce communication library
- Supports basic and advanced neural networks: MLP, CNN, RNN, GAN
- Exports to TensorFlow for finetuning, inference etc



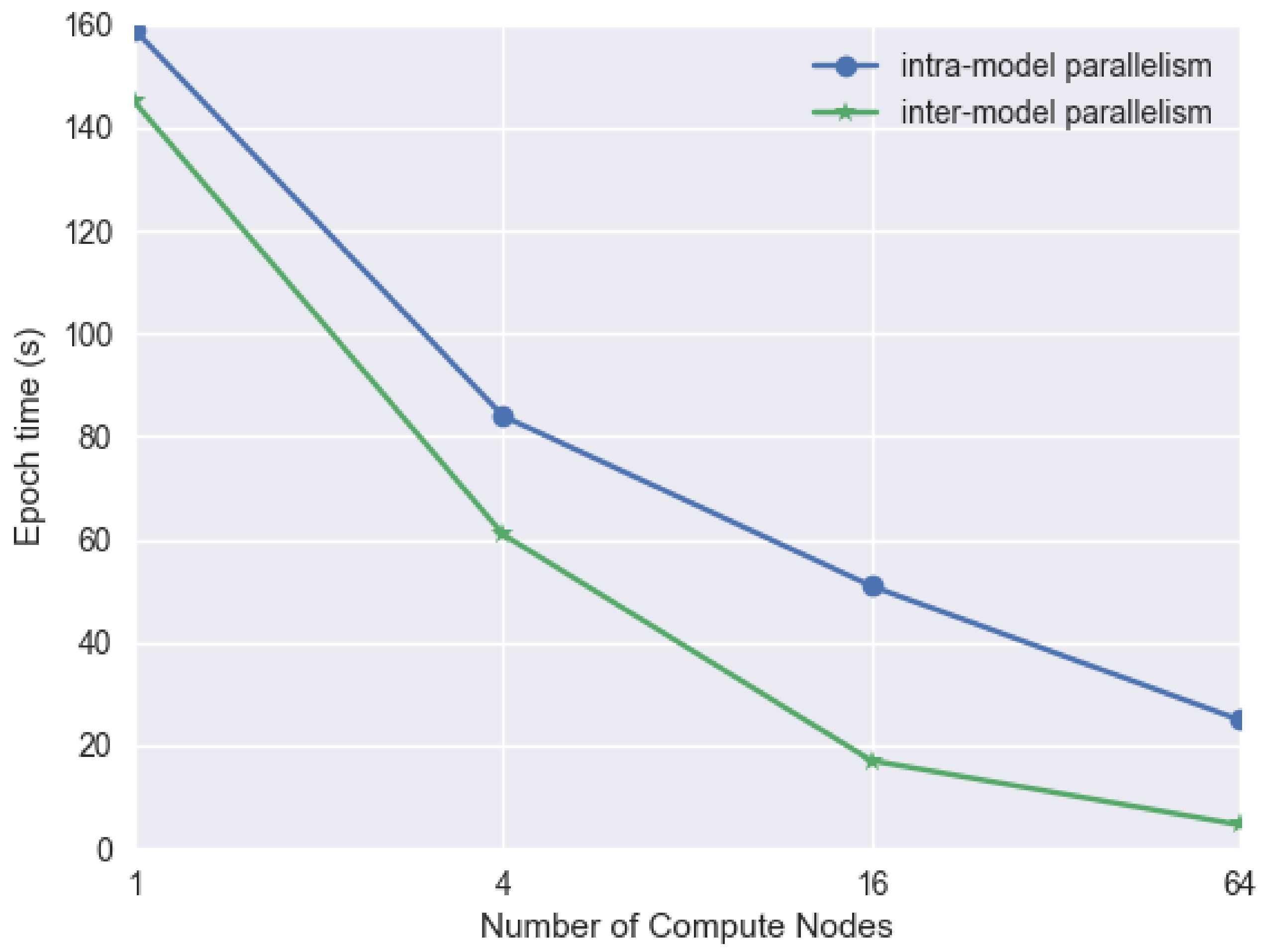
LBANN parallelism modes (distributing data samples and model parameters across processes)

Multiple Levels of Parallelism in LBANN

- Parallelism within one model instance (intra-model parallelism)
  - Model Parallel – distributed linear algebra
  - Data Parallel – distributed training data
- Parallelism across multiple model instances (inter-model parallelism)
  - Tightly coupled inter-trainer-peer-to-peer reduction trees
- Data parallelism across multiple models
  - Loosely-coupled inter-trainer data parallel tournament voting (LTFB[3])

EXPERIMENTAL RESULTS

- Dataset from NCI-ALMANAC database preprocessed for ingestion into LBANN. The preprocessed dataset consist of 244501 and 61125 training and test (validation) samples of cell-drugs combination respectively. Each sample consist of 8579 features
- Trained the Combo model on Catalyst HPC cluster at LLNL[4] with the following network architectures and hyper parameters: FC layers each of 1000 neurons, ReLU activations, dropout, Adam optimizer, 0.001 learning rate, 256 minibatch, and MSE loss



Strong scaling studies of Combo model

CONCLUSIONS

- Data-intensive scientific applications such as genomics present challenges and opportunities for exploring deep learning at scale
- We demonstrate LBANN as a toolkit for addressing problems of large data, large models training with promising results for further exploration (e.g., extension to complex models that predict the effect of multiple drugs on a tumor)

REFERENCES

1. Fangfang X. et.al., Predicting Tumor Cell Line Response to Drug Pairs with Deep Learning, CAFCW17
2. https://github.com/llnl/lbann
3. Jacobs S.A., Dryden N., Pearce R., Van Essen B., Towards Scalable Parallel Training of Deep Learning, MLHPC17
4. https://hpc.llnl.gov/hardware/platforms/catalyst

Towards Scalable Machine Learning for Scientific Discovery in HPC Environments

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.