Nested Active Learning for Efficient Model Contextualization and Parameterization

The University of Vermont
LARNER COLLEGE OF MEDICINE

Chase Cockrell and Gary An, Department of Surgery, University of Vermont

Background - Sepsis

- Dysregulation of inflammatory signaling network dynamics
- Affects ~1 million people/year
- Mortality rate: 28-50%
- Treatments:
  - Focused on manipulating single mediator/cytokine
  - Single dose or very short course (<72 hrs)
- Reasons for failure:
  - Nonlinear inflammatory signaling network
  - Chaotic "error" propagation due to individual response

Background – Modeling Philosophy

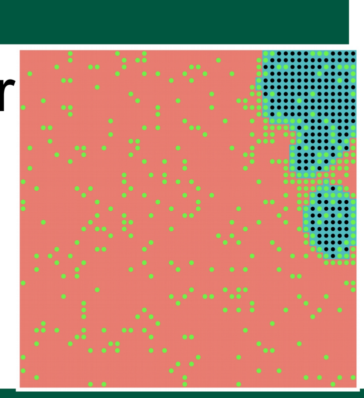
- Model Content: parameterization of internal model rules
- Model Context: description of the environment in which a biomedical simulation operates
- Defining the boundaries of model content and context is necessary to represent biological heterogeneity in complex dynamical models

Nested AL Workflow Pseudocode

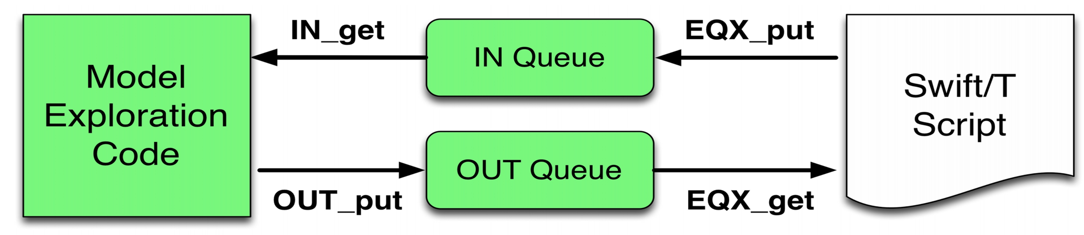
1. Initialization of initial dataset *I* (consisting of the Internal Parameterizations), training pool *P*, number of samples added on each step *m*, the final size of the dataset *f*, network architecture, and learning parameters.
2. Train network on *I*.
3. While |*I*|<*f*:
   a. obtain the rank *rj* for every *xj in P* using an acquisition function, *A*
   b. Sample point set *S*; |*S*|=*m* with maximal variance ranks r_j
      i. Perform AL to determine boundaries of CR space using External Parameterization dataset
      ii. Return volume, center-point of CR space
   c. Add *S* to *I*
   d. Remove *S* from *P*
   e. Train NN on *I*
   f. Calculate stopping metrics, stop if appropriate

AL Visualization Example

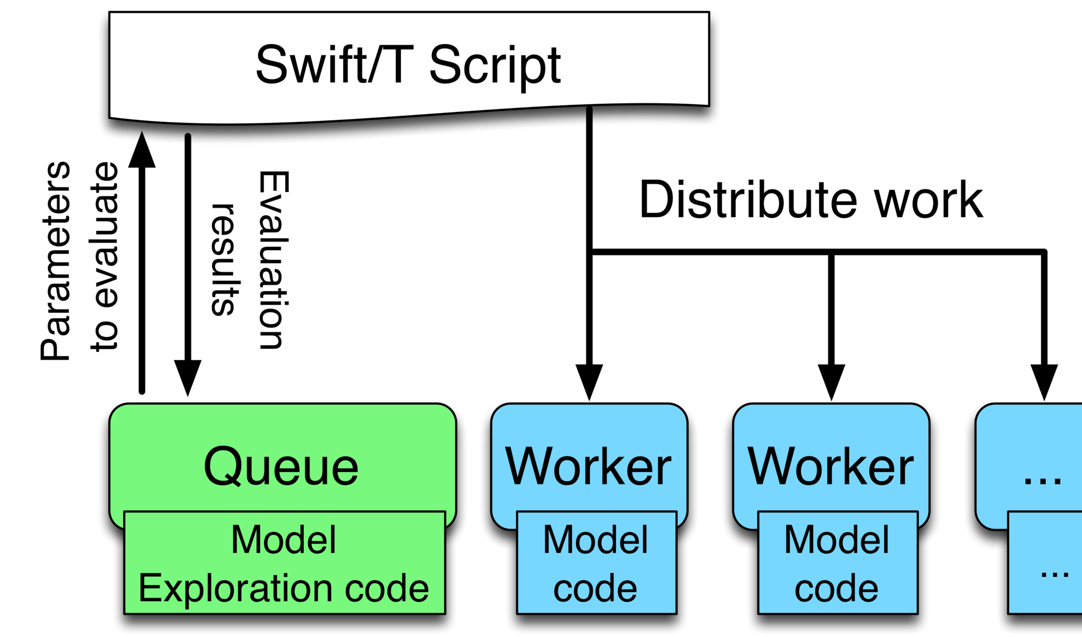
In order to test the generalizability of our lower-level AL scheme, we tested on a variety of synthetic data. Red: class 1; Teal: class 2; Green: sampled pts; Dark Blue: predictions for class 2



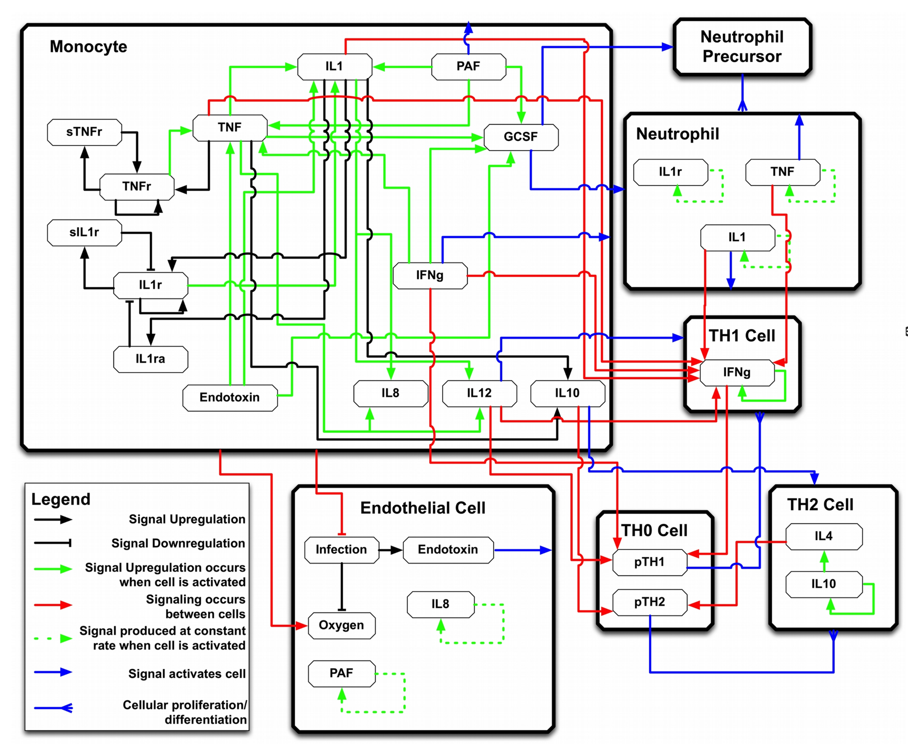
Extreme-scale Model Exploration With Swift (EMEWS) Workflow

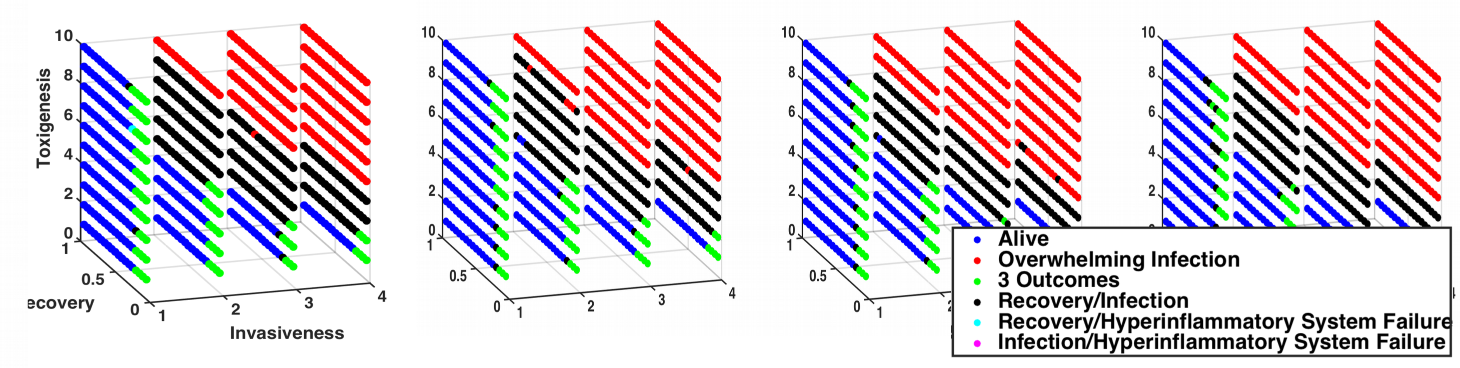


EMEWS combines existing machine learning/model exploration libraries (i.e., Keras, Tensorflow) with the **Swift/T** parallel scripting language to run scientific workflows in an HPC environment. This work was performed on the Edison supercomputer at NERSC.

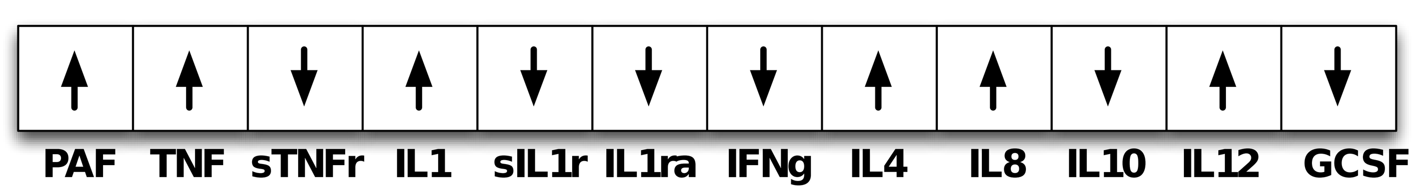


Model and Methods

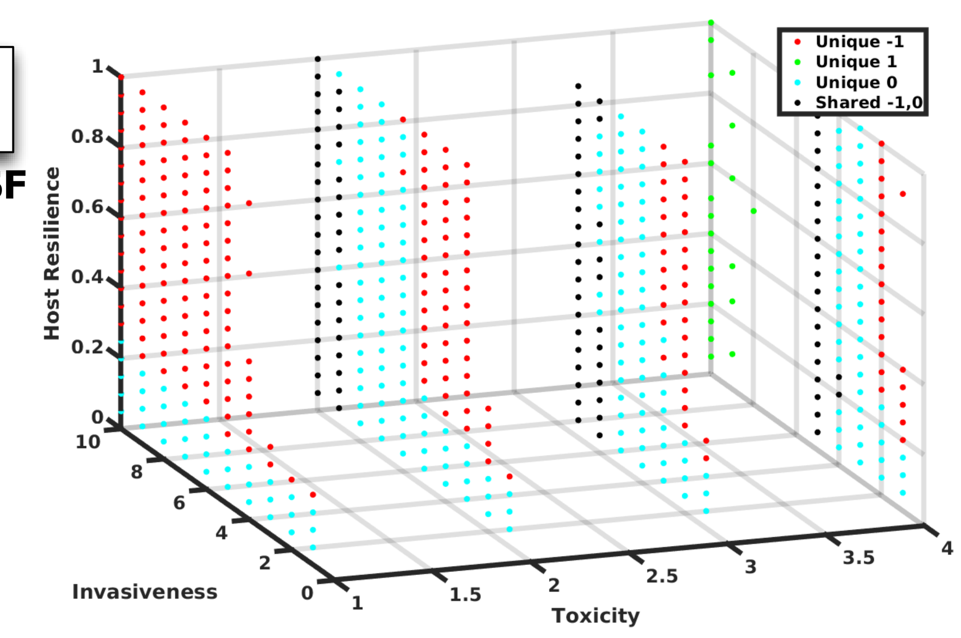




Model Content is shown on the left as an influence diagram between key elements in the simulation. Model context is illustrated above and represents the environmental circumstances under which an *in silico* patient develops sepsis.

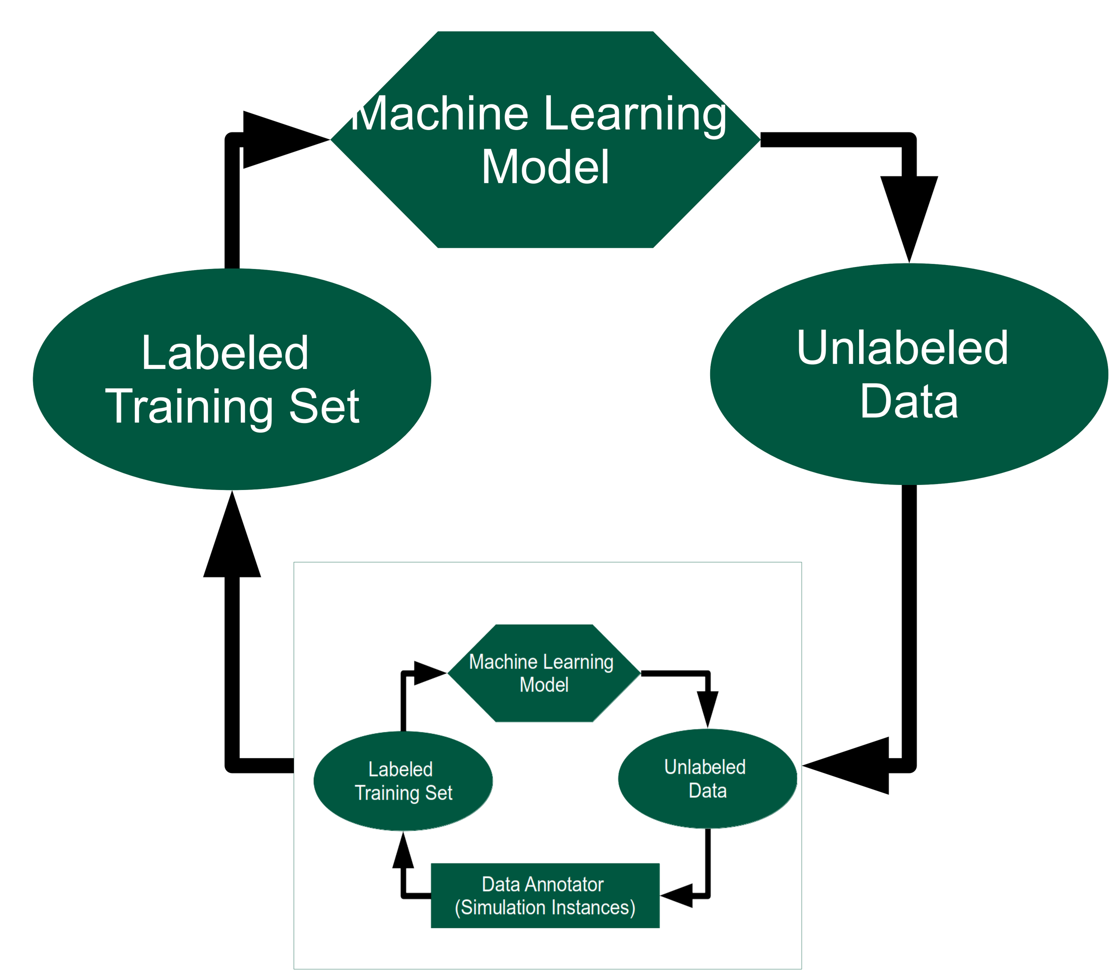


Genomic variability is simulated via augmentation or inhibition of key elements in the model's internal cell signaling network. Different internal parameterizations can lead to vastly different CR spaces

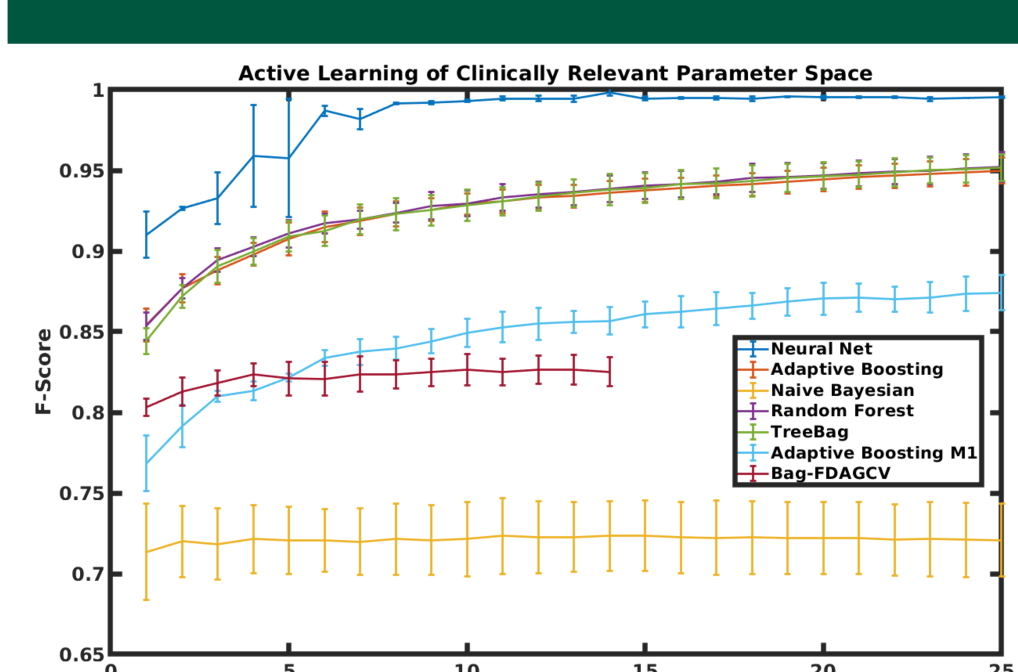


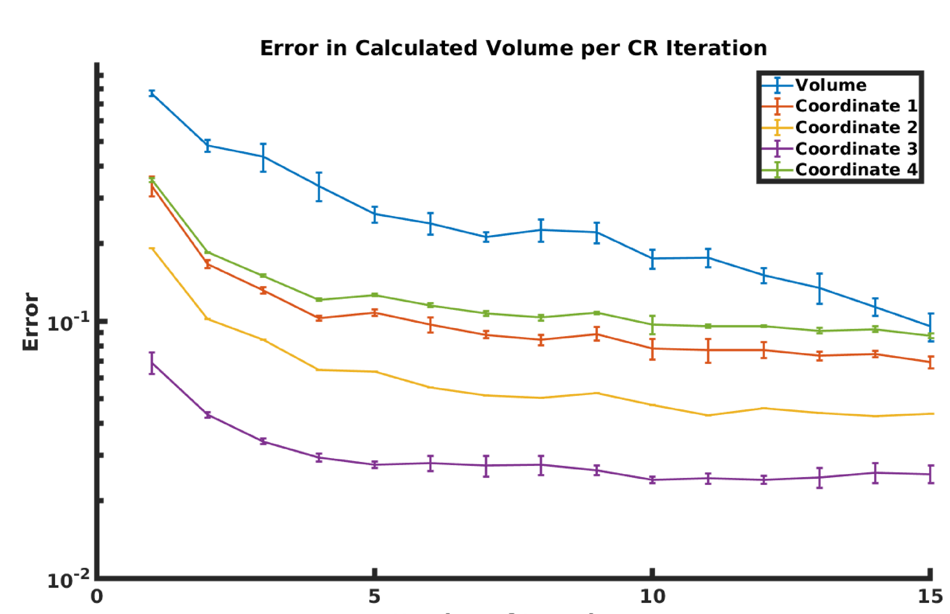
Active Learning:

- Used when there is lots of unlabeled data, data expensive to label
- Algorithm adaptively queries data
- Lower level AL determines the clinically relevant region of parameter space for a given internal parameterization
- Lower-level AL seeks to minimize uncertainty in class prediction (clinically relevant or not)
- Upper-level AL predicts CR volume and centroid location.
- Upper-level AL samples parameterizations which maximize output variance



Results





We tested a variety of machine learning models to be used in the lower-level AL module. Artificial Neural Networks were the superior option for this problem, both in terms of simulation accuracy and efficiency, and in terms of total wall-time necessary to complete the calculation. The Upper-Level AL scheme determines the centroid location with very minimal error; error in predicted volume appears to stabilize slightly above 90%.

The lower-level AL achieves >95% accuracy while sampling an average of 2% of the possible external parameterizations. The upper-level AL regression stabilized after seeing approximately 2000 samples out of over 40 million evenly discretized internal parameterizations. Using nested active learning instances, **we have generated comprehensive map linking model content and context using only 1/1,000,000 of the simulations that would have been required using brute-force.** We anticipate that more advanced techniques will lead to greater gains in both efficiency, accuracy, and utility.

Acknowledgments

This work was supported by MSM U01 Grant Number: **EB025825-01**. Additionally, this research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231.