Enabling Multiscale Simulations of RAS Biology using Machine Learning

Harsh Bhatia
Center for Applied Scientific Computing

P Karande, G Dharuman, C McNeish, R Berg, H I Ingolfsson, T Carpenter, L Stanton, J Glosli, T Oppelstrup, F Lightstone, B V Essen, F Streitz, P-T Bremer

Data Science Institute Workshop
Aug 07, 2018

Lawrence Livermore National Laboratory

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory (LLNL) under contract DE-AC52-07NA27344. LLNL-PRES-755190

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

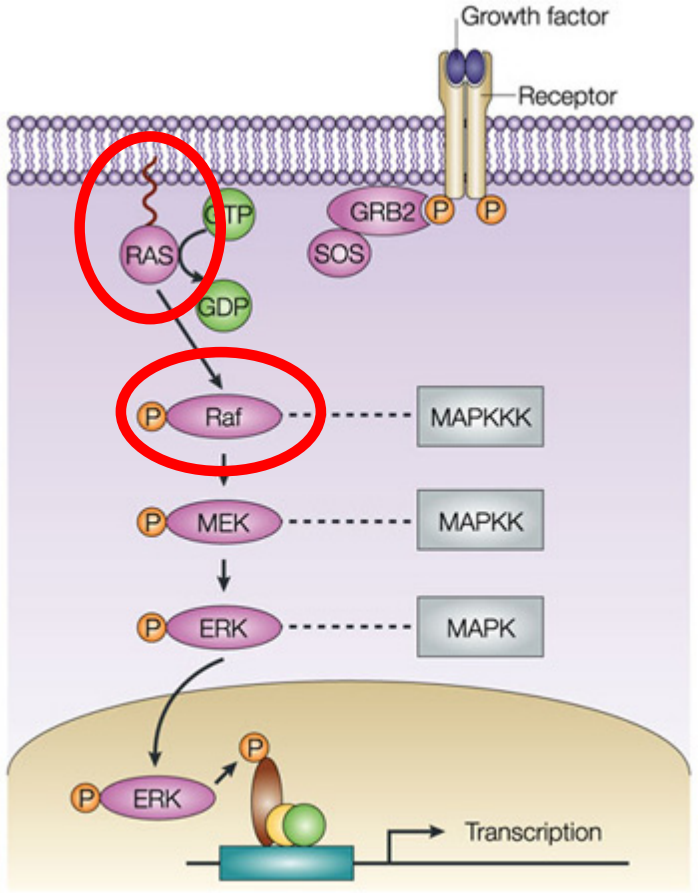
Pilot Project 2: RAS-Related Cancers

NCI-DOE Collaboration

(LLNL, LANL, ORNL, & FNLCR)

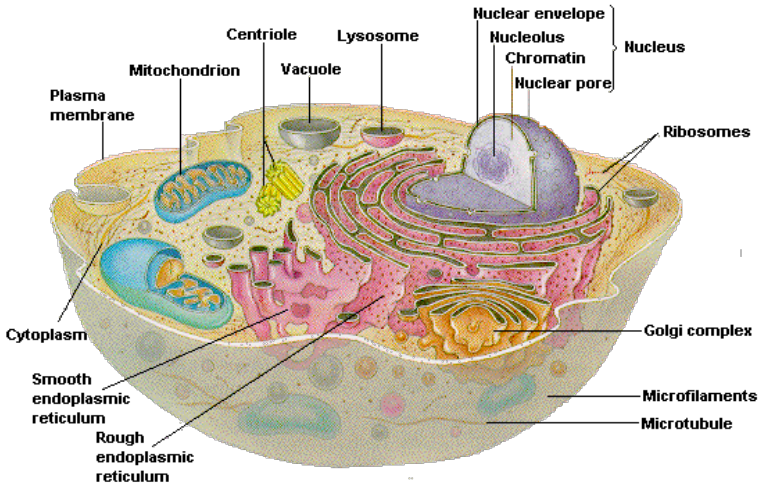
- Mutated RAS protein is responsible for 1/3 of all human cancers, yet remains untargeted with known drugs
  - **95%** of all pancreatic; **45%** of all colorectal; **35%** of all lung cancers
  - **>1 million** deaths/year
  - **No** effective inhibitors

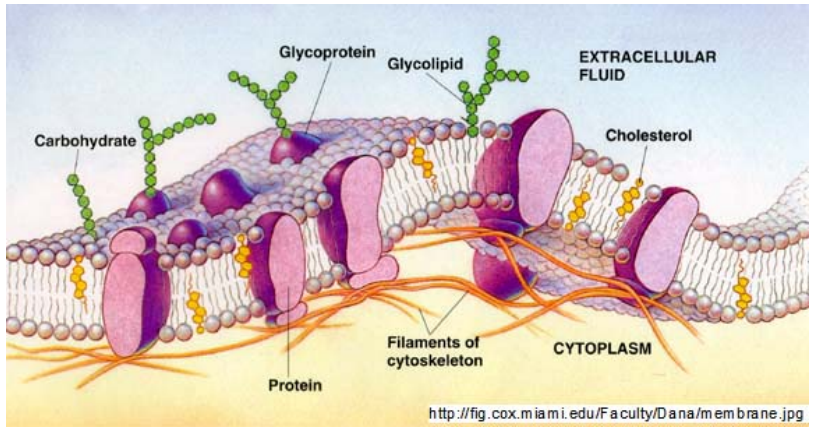
- New insights are required to develop better models
  - Simulations and predictive models for RAS-related molecular species and interactions
  - Advanced multimodality data integration

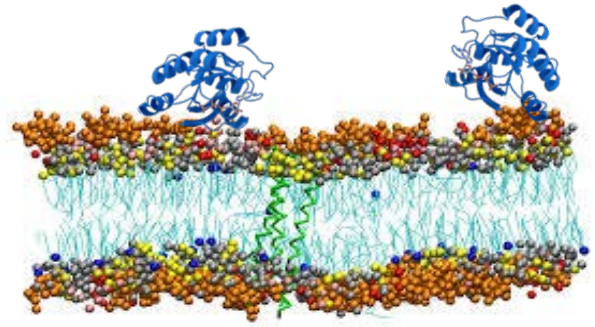


Nature Reviews | Molecular Cell Biology

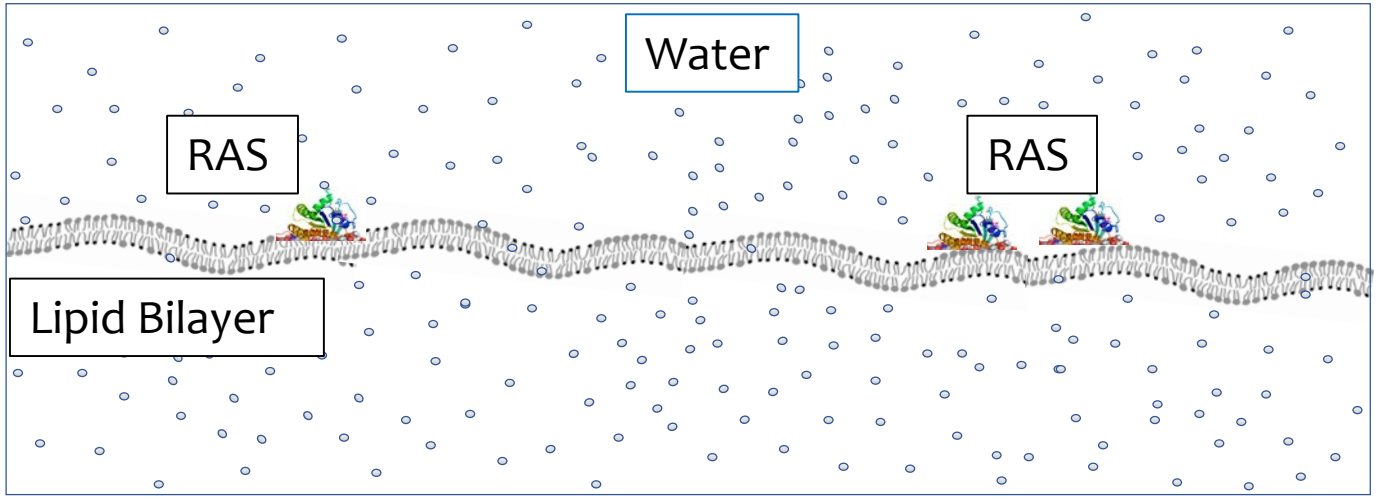
Understanding RAS biology requires modeling at different scales

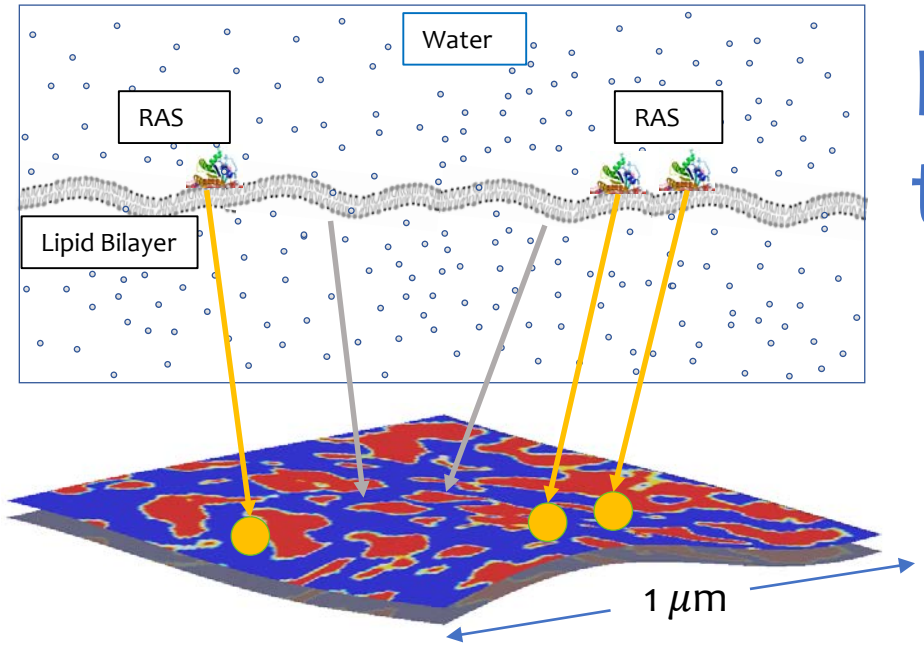




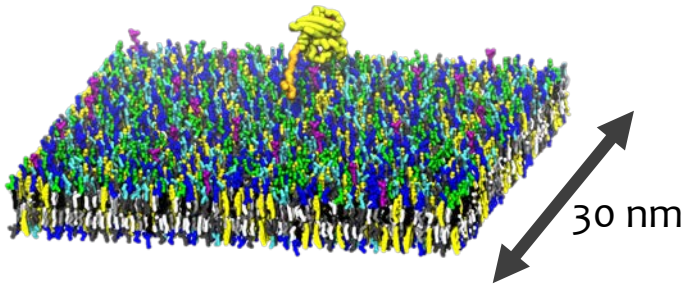


- RAS-RAS ~1 ms
- Lipid flips ~1 ms
- RAS diffusion ~200 ns
- Lipid diffusion ~40 ns
- Bond vibration ~2 fs





Different scales require different types of simulation



(DDFT) Macro Model

- Lipid bilayer and water represented by Phase Field (PF)
- Proteins represented by a Hyper-Coarsen Particle (HyCoP)
- PF+HyCop parameters determined via simulations and experiment

- **One** macro simulation

- 1 ms per day
- Domain size: ~[1 μm x 1 μm]
- Time step: ~0.05 μs

(Coarse grain) Micro Model

- Coarse grain (CG) bead model using the Martini force field
- Mapping of ~10 atoms to a single bead

- **~130K** bead simulations

- 1 μs per day (per simulation)
- Domain size: ~[30 nm x 30 nm]
- Time step: ~ 20 fs

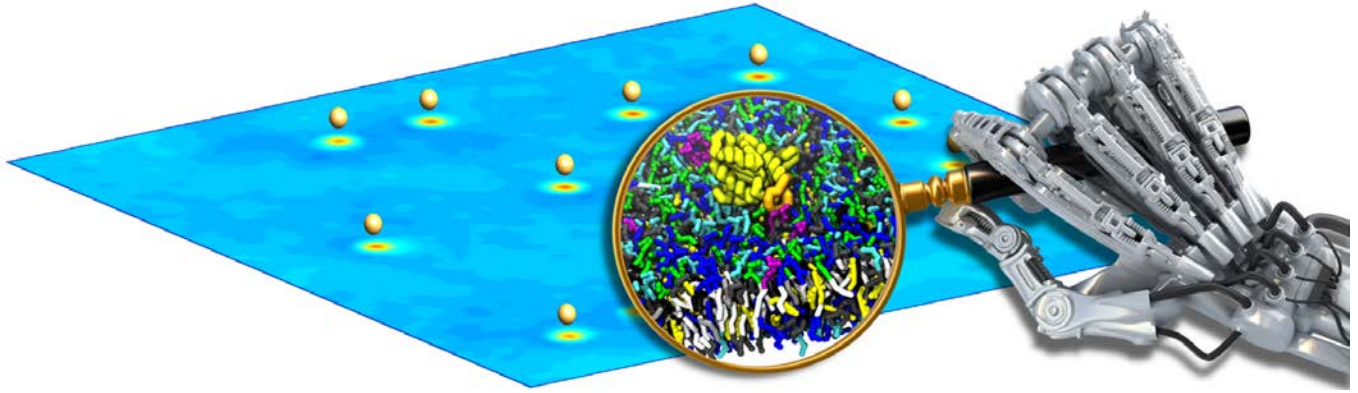
(Atomistic) Micro Model

- All atoms (AA) simulations using the CHARMM force field

- Ensembles of **~1.5 million** atoms simulations

- 1 μs per day (per simulation)
- Domain size: ~[30 nm x 30 nm]
- Time step: ~ 2 fs

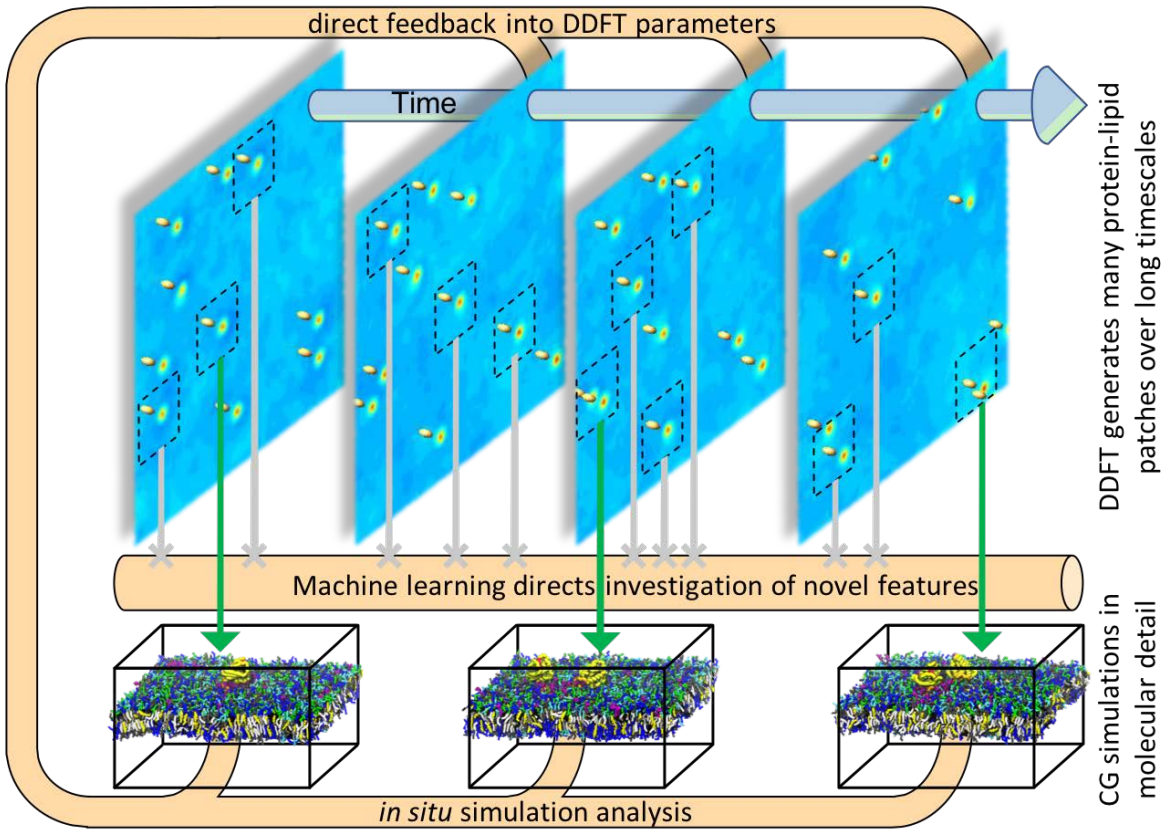
Enabling the *virtual microscope* of multiscale simulations via Machine Learning



- When and where to use a resolve to finer scale?
  - "*Interesting*" configurations

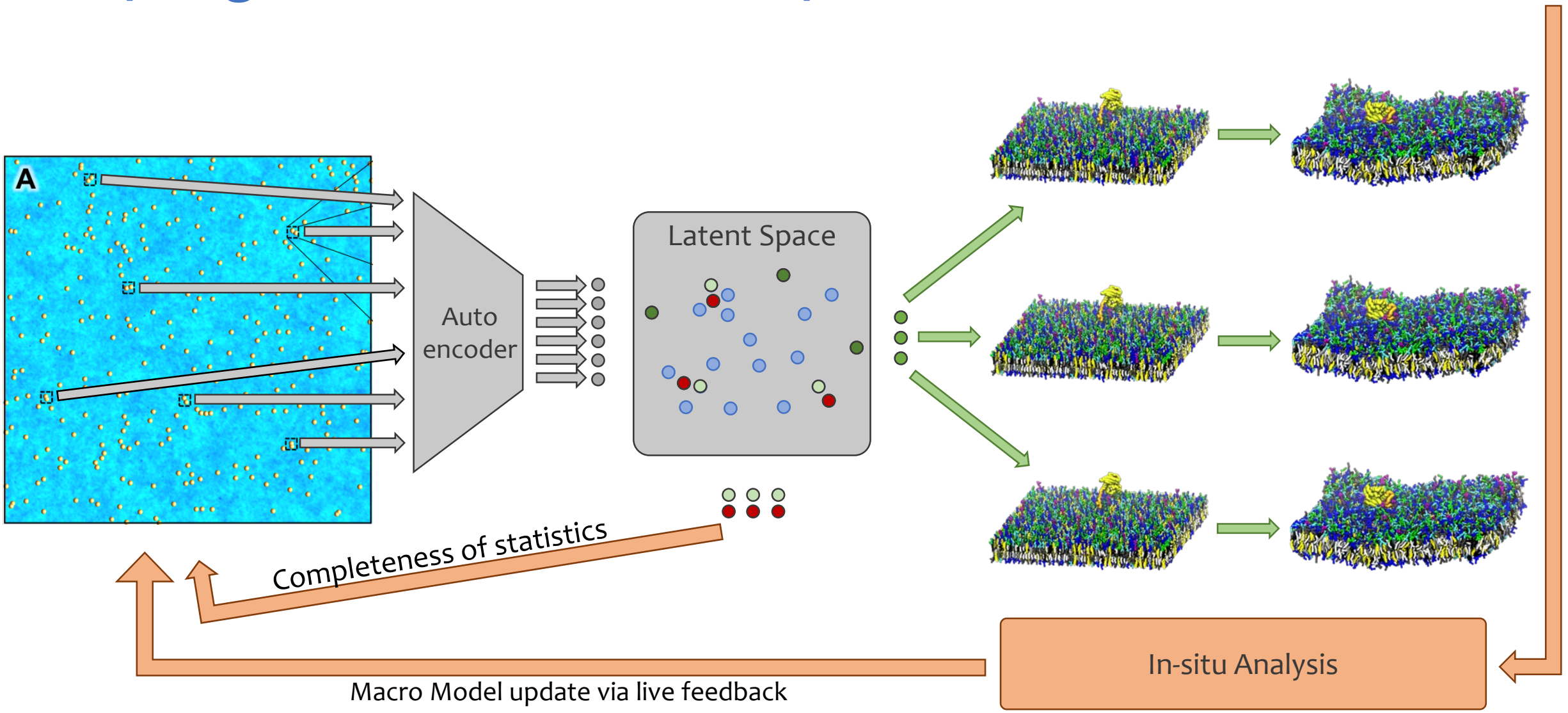
- Is the multiscale simulation self-consistent?
  - Consistency in heterogenous data from different simulations

Current work focuses on coupling between Macro and CG Micro model



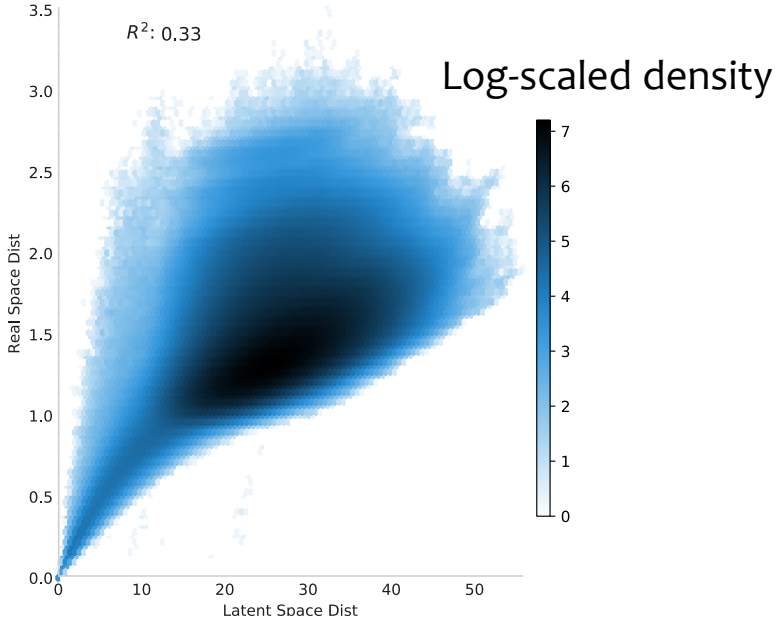
- Macro model simulation
  - ~1000 nodes
  - Running for several days
- Machine learning
  - Continuously ranking potential candidates
- CG simulations and analysis
  - Top candidates scheduled as and when resources become available
  - ~3500 nodes
  - running for several hours

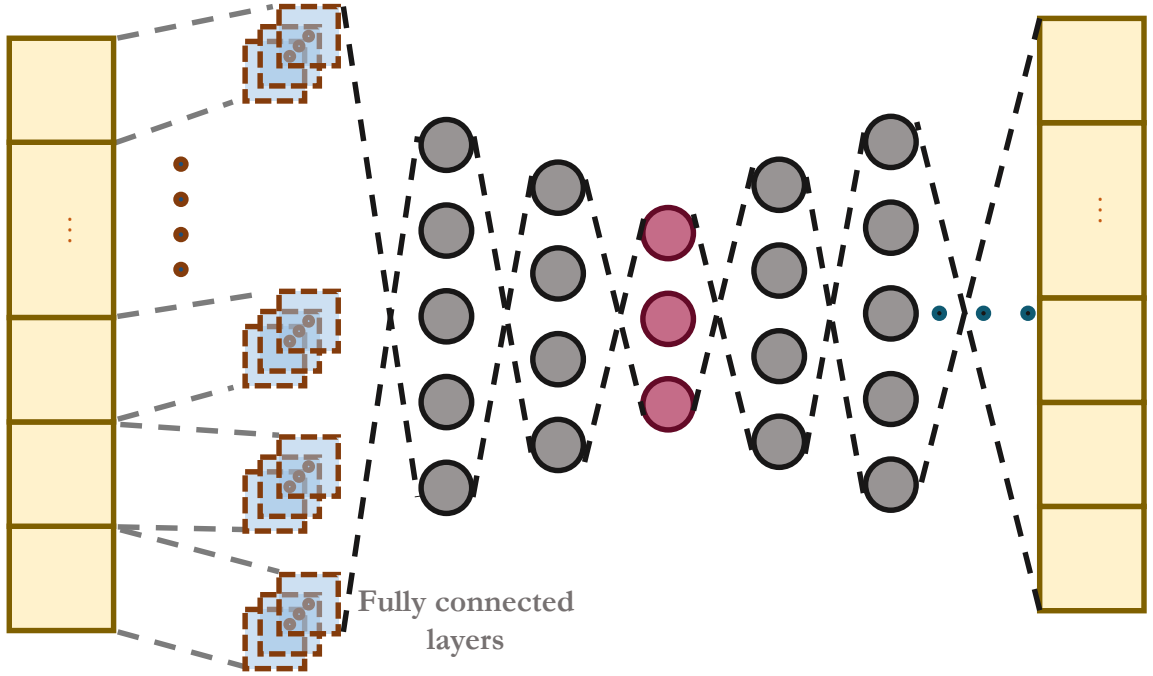
Steering the multi-scale simulations by adaptively sampling data-driven latent space



We need a specially designed model to suit the data of interest

- Optimized for a custom ℓ_2 reconstruction loss
  - Convolutional layers
  - Fully-connected layers
  - Batch-normalization layers
- Preserve relative ℓ_2 distances





Input data: 5x5x14 grid

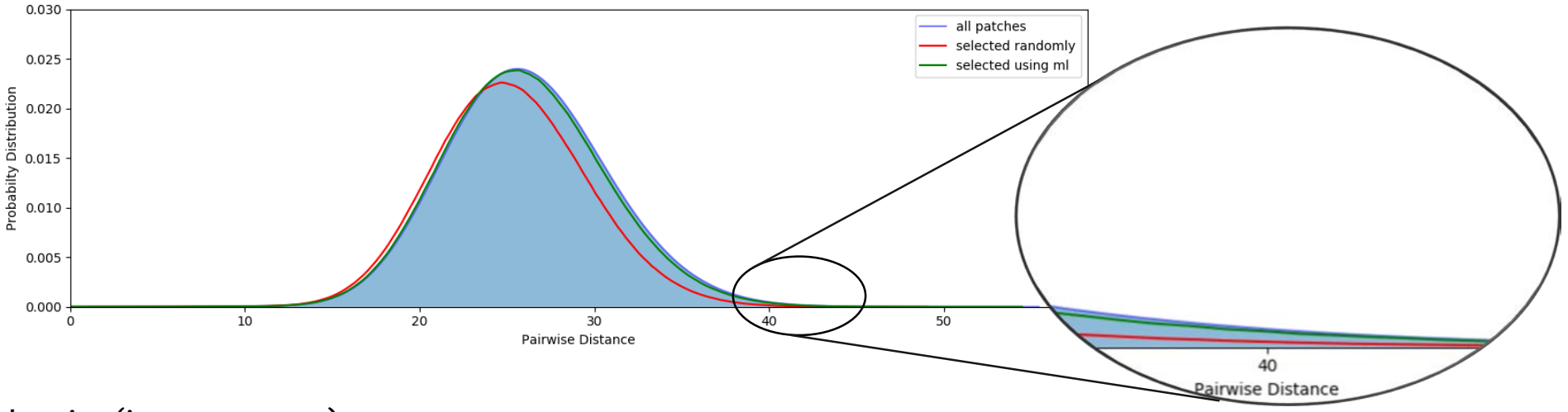
1D Convolutional filters: 25x256 for capturing across-channel variation

Compressed latent space representation: 20D

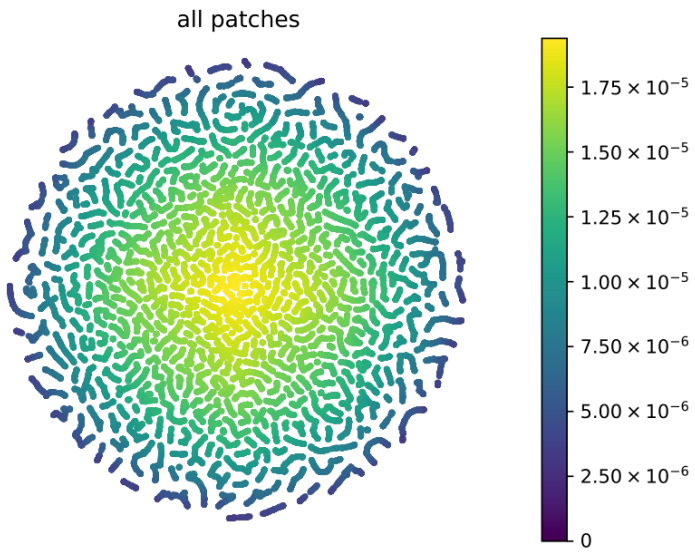
Reconstruction

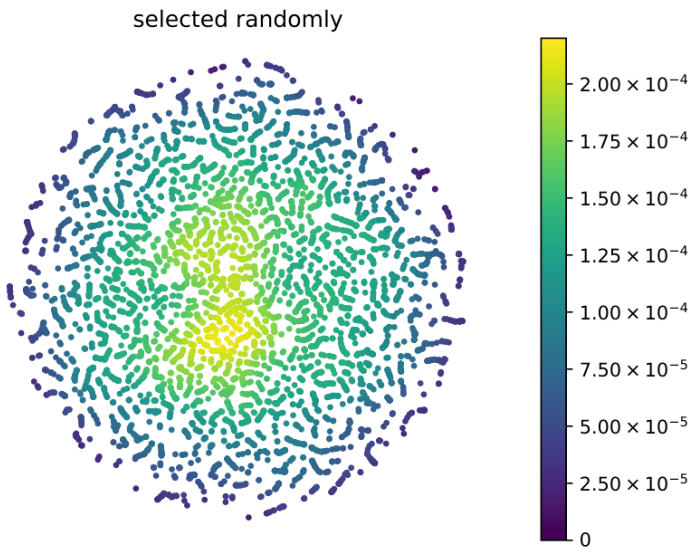
Our selection criterion picks rare events

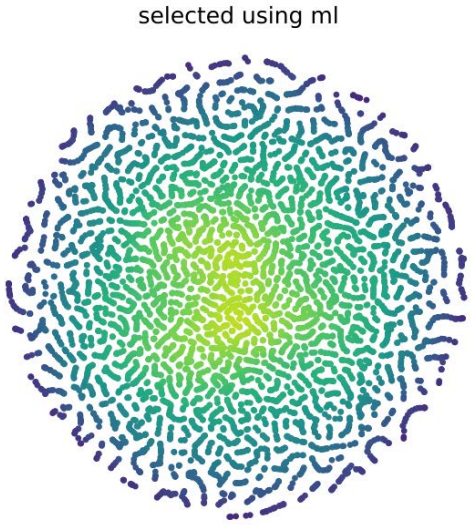
distribution of
pairwise distances



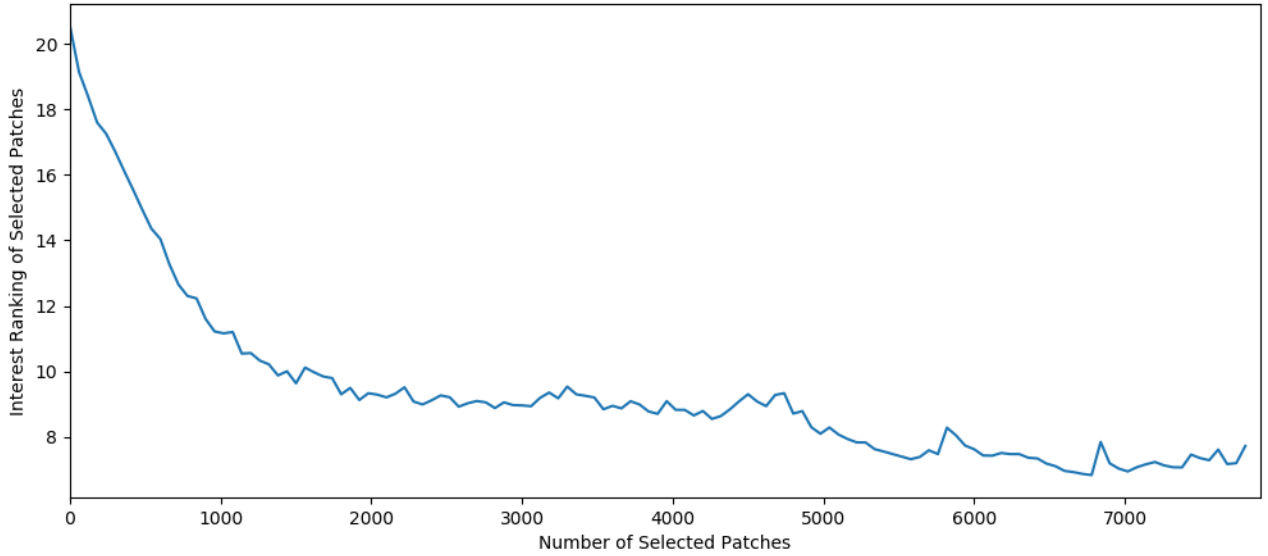
t-sne embedding colored by density (in t-sne space)







With progressing multiscale simulation, we expect to cover all regions of interest



By the end, we would have run the simulation effectively at micro scale, but using macro model

What's next?

- Properties of ml sampling further need to be explored
- Sampling should also incorporate other parameters, e.g., RAS state
- Coupling between CG and AA models needs to be established
- Online training of the model
- More complex types of feedback