

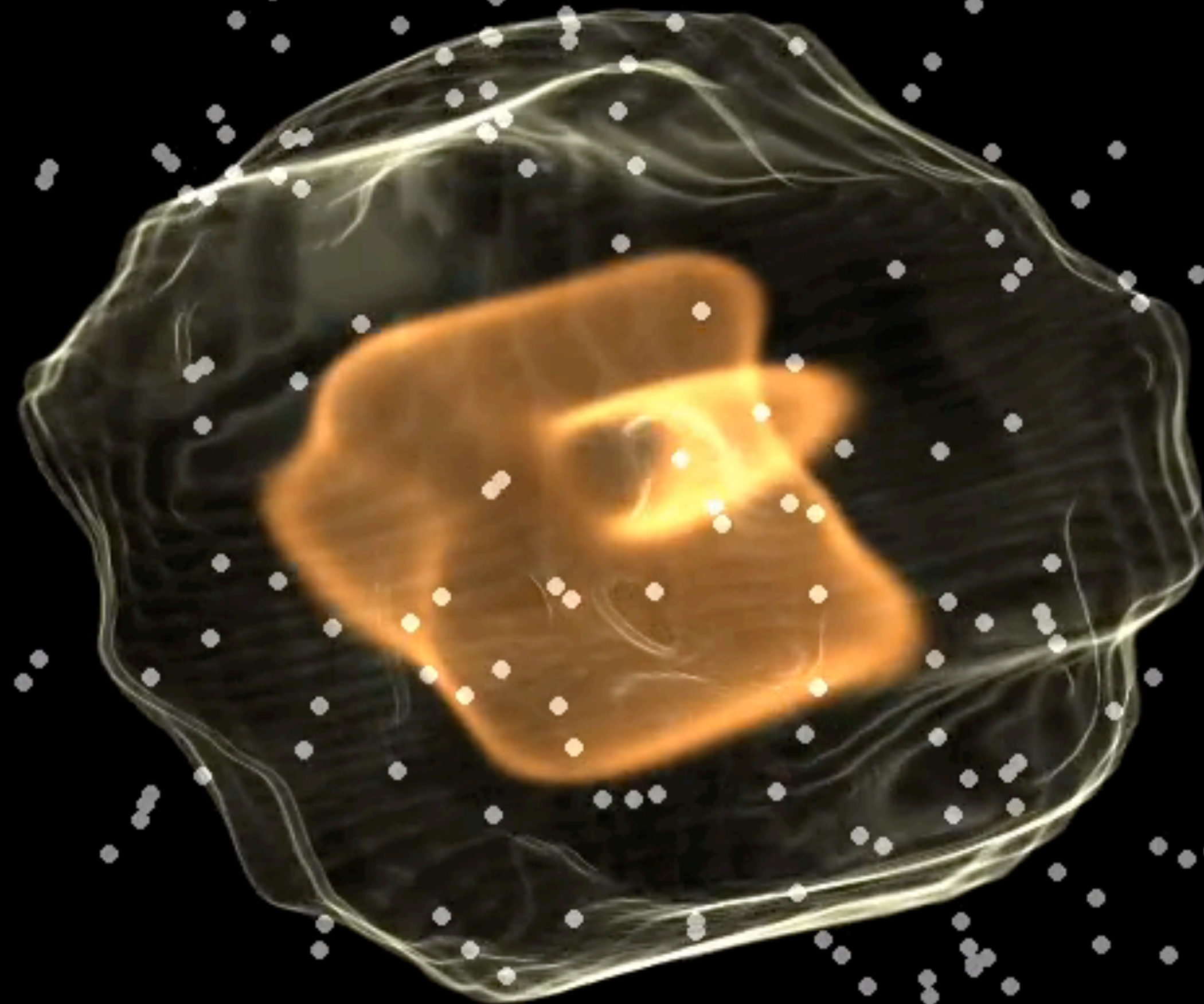


OPEN DATA FOR NP SIMULATIONS

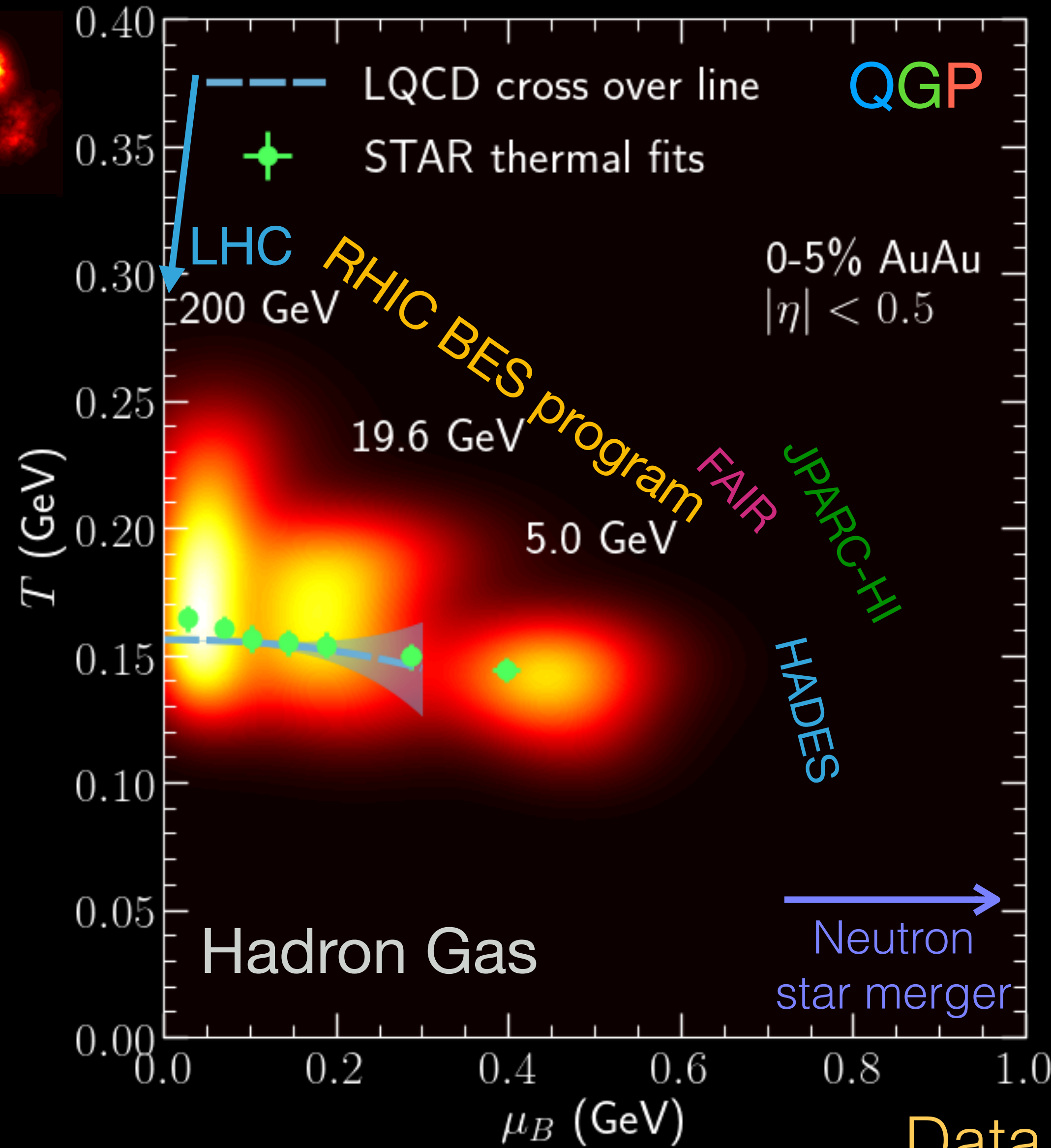
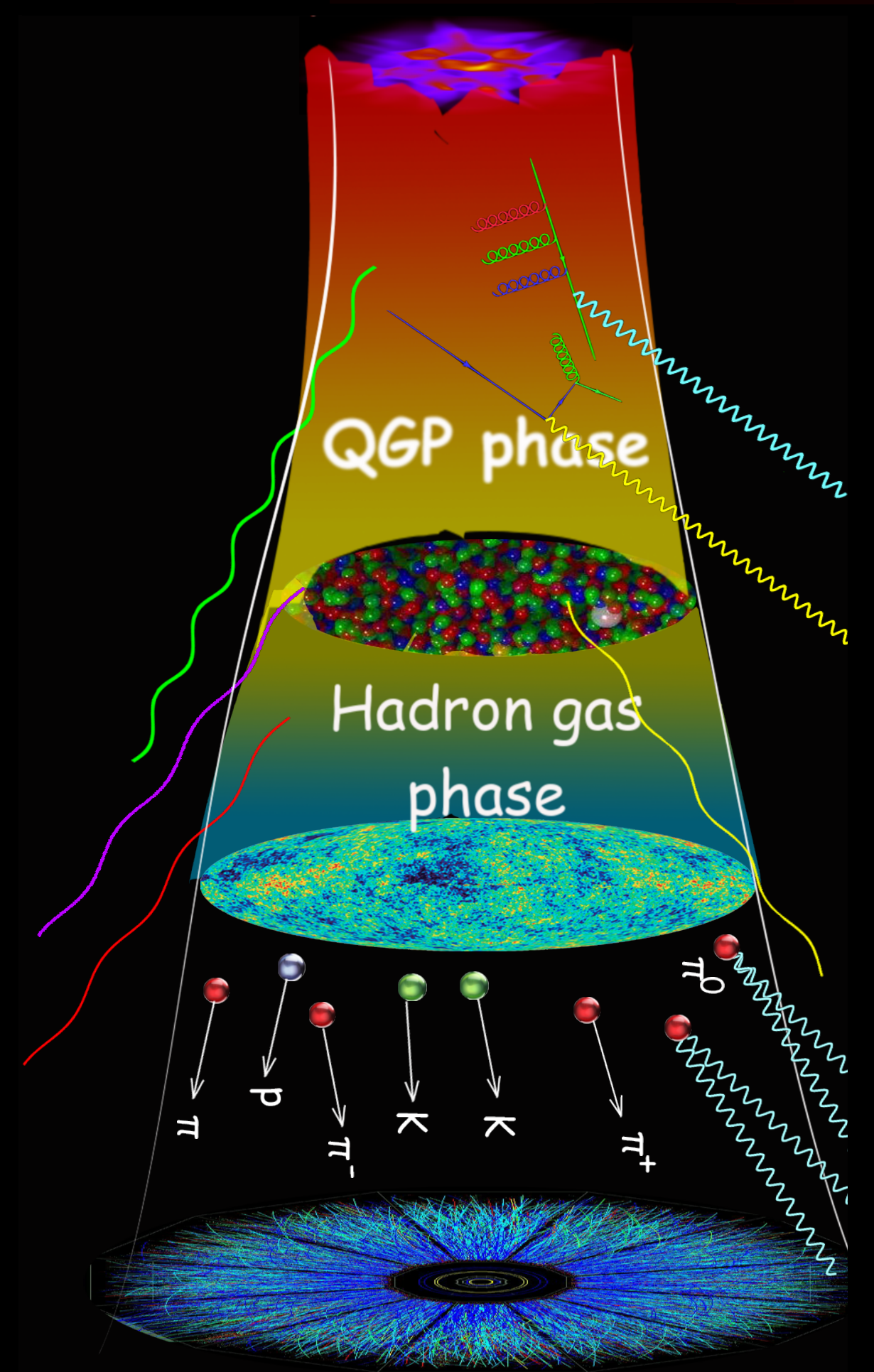
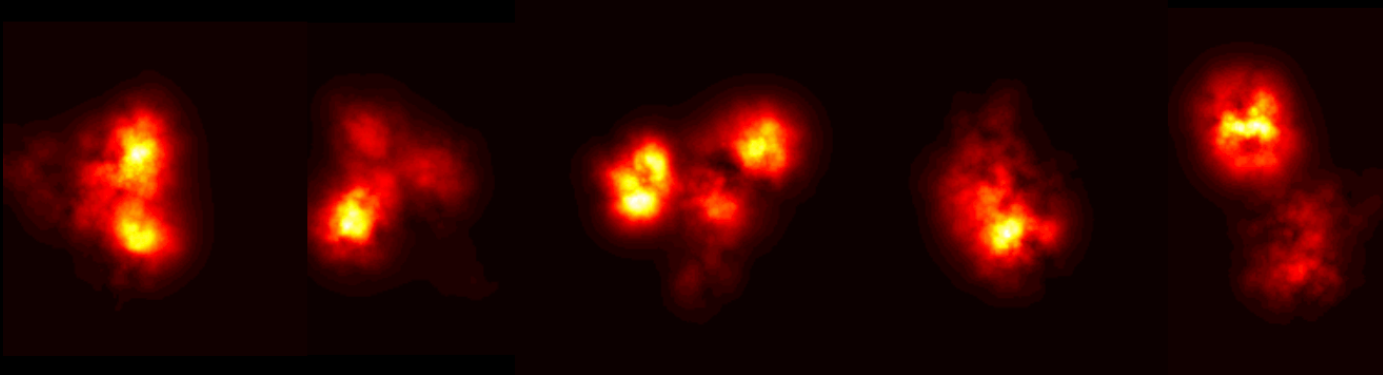
CHUN SHEN

SANPC 2024, Jefferson Lab

June 20, 2024



NUCLEAR MATTER UNDER EXTREME CONDITIONS



- Simulating relativistic heavy-ion collisions is computational intensive

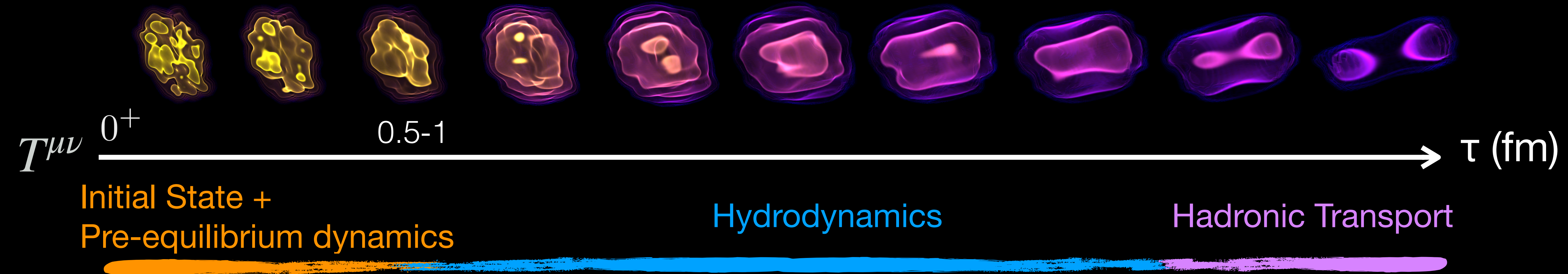
$O(10)$ hours per event

- High volume simulations ($O(10^6)$ events) are required for Machine-learning, Bayesian inference, Quantitative comparisons with measurements

Data generation is expensive!

WHY OPEN DATA IS CRUCIAL FOR ML IN HIC?

- Dynamical description of relativistic heavy-ion collisions is *multi-stage*,



- ML and Bayesian analyses require high volume of simulation data with various model combinations, which exceeds a single research group's capacity
- Having data open for community usage from multiple groups can significantly boost ML activities in our field

OPEN DATA WITH MULTIPLE HIERARCHY

- Data storage is expensive
- Zero: A tagged version of code container + parameter + Open Science Grid
- Light $O(GB)$: Sharing event-averaged observables at training design points
 - Ideal for model emulation in Bayesian Inference analysis
- Moderate $O(TB)$: Particle yields and flow vectors as functions of p_T and η for every simulation event
 - Ideal for general model emulation at event-by-event level
 - An effective way to communicate with experimentalists on new analyses
 - Require a few iterations with community users
- Full $O(PB) - O(EB)$: Final-state hadrons momentum information for the full event; end-to-end event evolution history