Nuclear matter under extreme conditions and High-Performance Computing

Gojko Vujanovic

University of Regina

CINP Town-Hall
CAP Congress 2025
University of Saskatchewan

Saskatoon, Saskatchewan

June 13th, 2025

Canadian Association of Physicists
Association canadienne des physiciens et physiciennes

University of Regina

Canada Research Chairs
Chaires de recherche du Canada

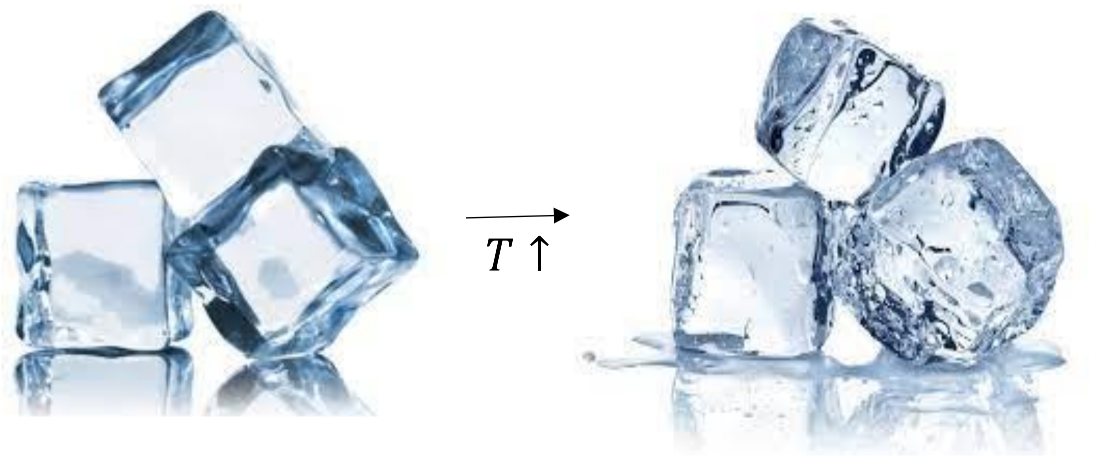
Natural Sciences and Engineering Research Council of Canada
Conseil de recherches en sciences naturelles et en génie du Canada

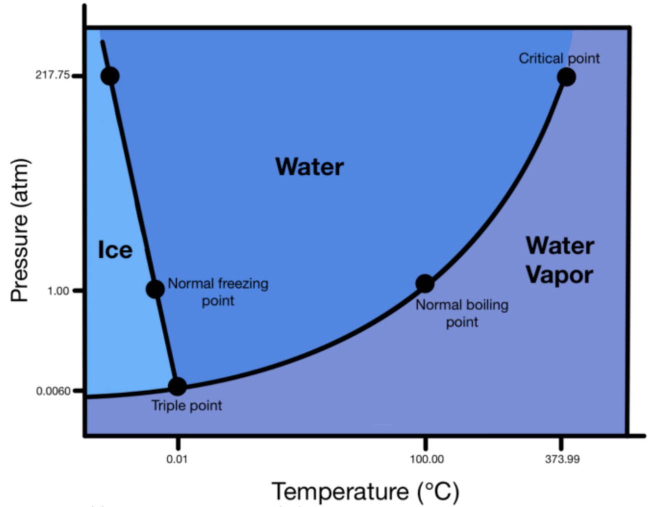
Canada

innovation Saskatchewan

INNOVATION
Canada Foundation for Innovation
Fondation canadienne pour l'innovation

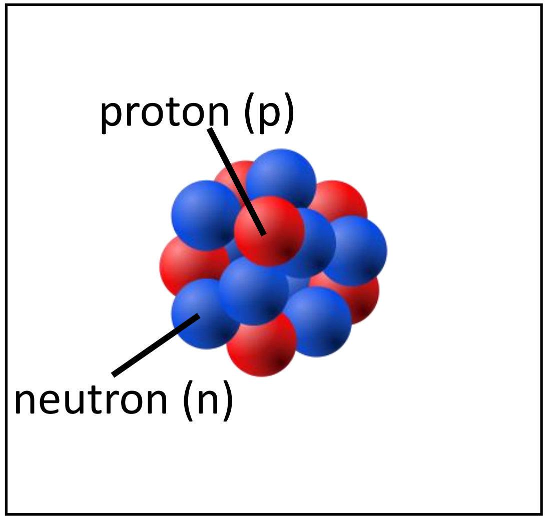
High energy nuclear collisions & nuclear equation of state





Ref.: https://www.expii.com/t/phase-change-diagram-of-water-overview-importance-8031

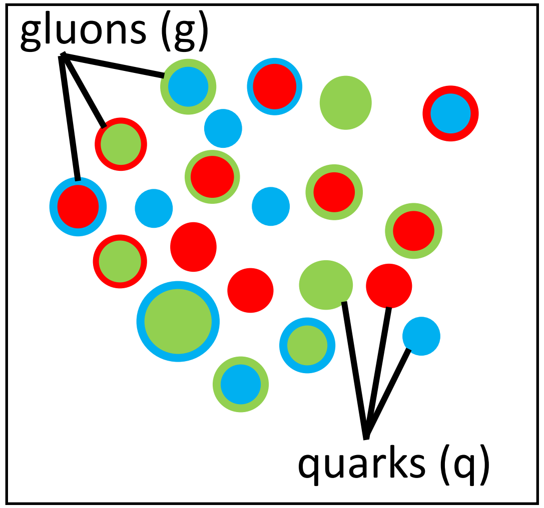
Ref.: https://en.wikipedia.org/wiki/Atomic_nucleus



$T\sim10^{2}\ Kelvin\ (10^{-2}\ eV)$

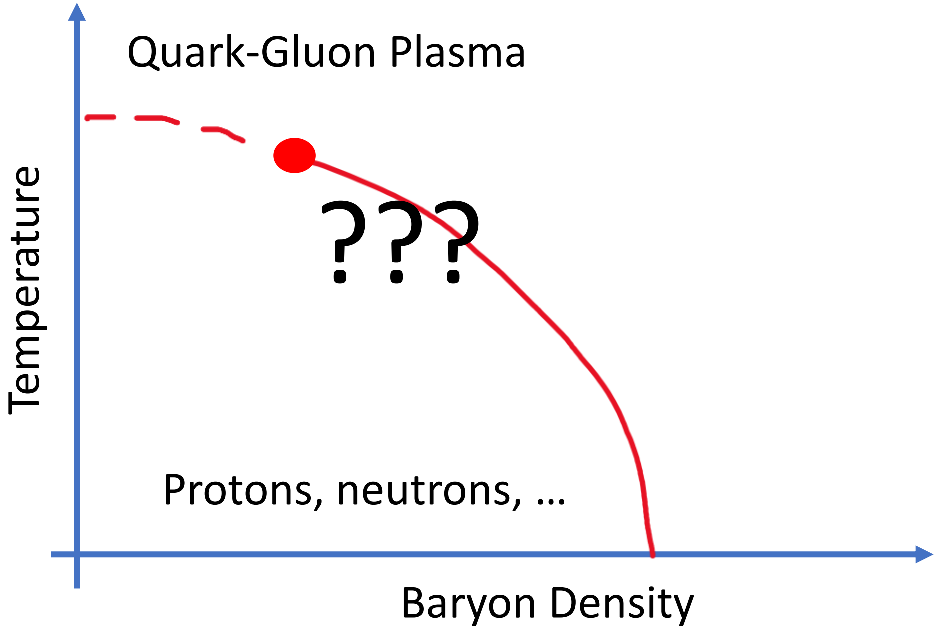
nucleus

$T \uparrow$



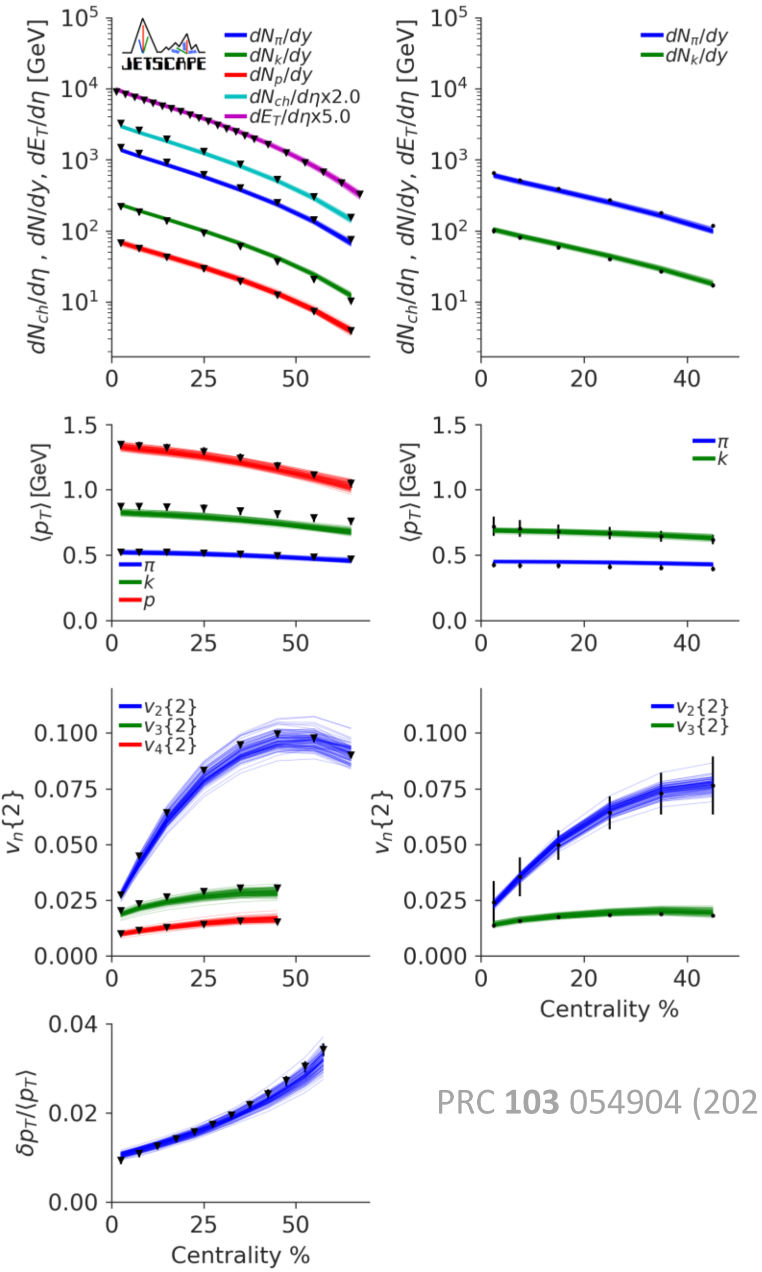
$T\sim10^{12}\ Kelvin\ (\sim 10^{8}\ eV)$

Quark Gluons Plasma



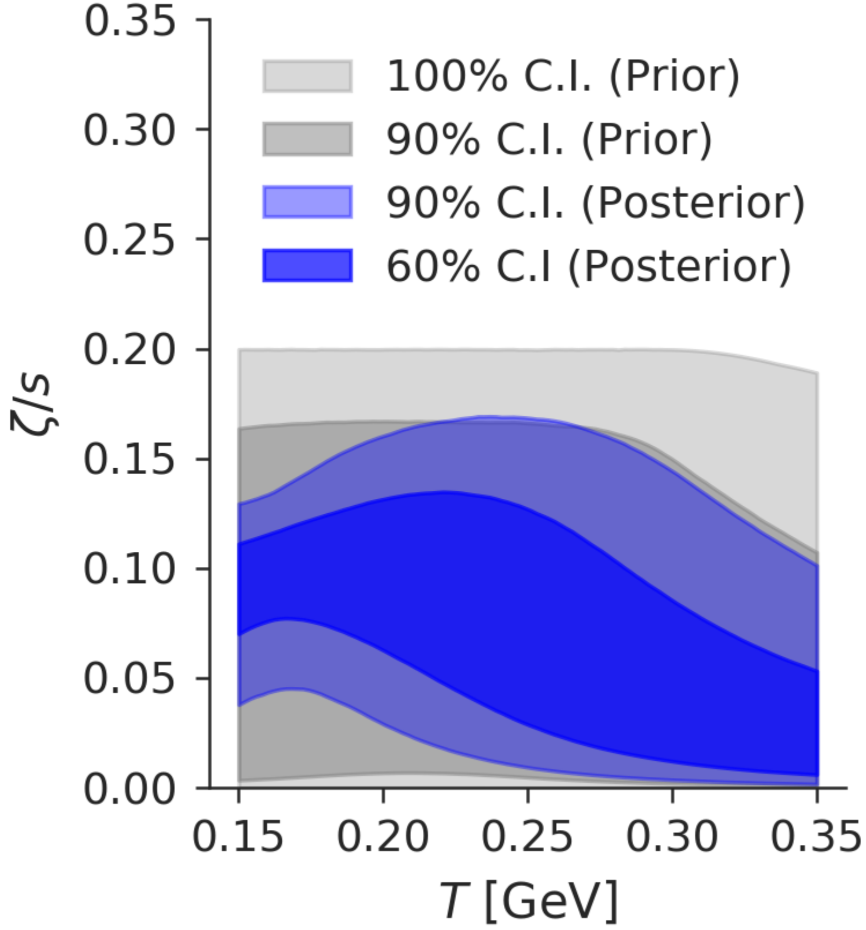
Comparisons w/ experimental data using Bayesian calibration

Observables Posterior : Grad

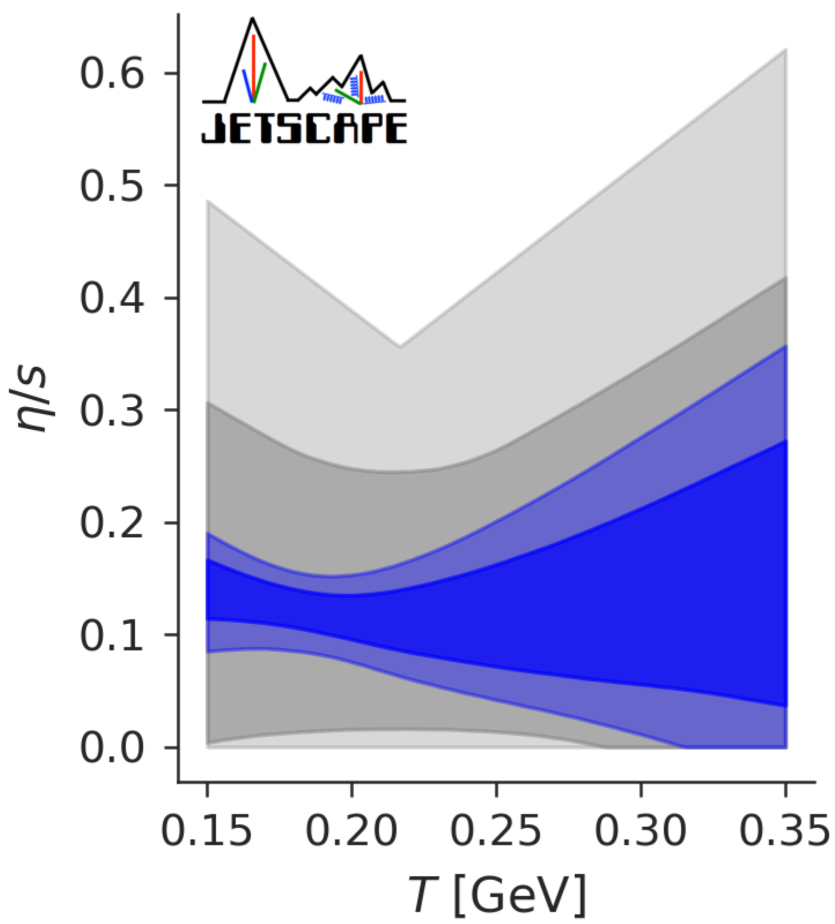


PRC **103** 054904 (2021)

PRC **103** 054904 (2021)



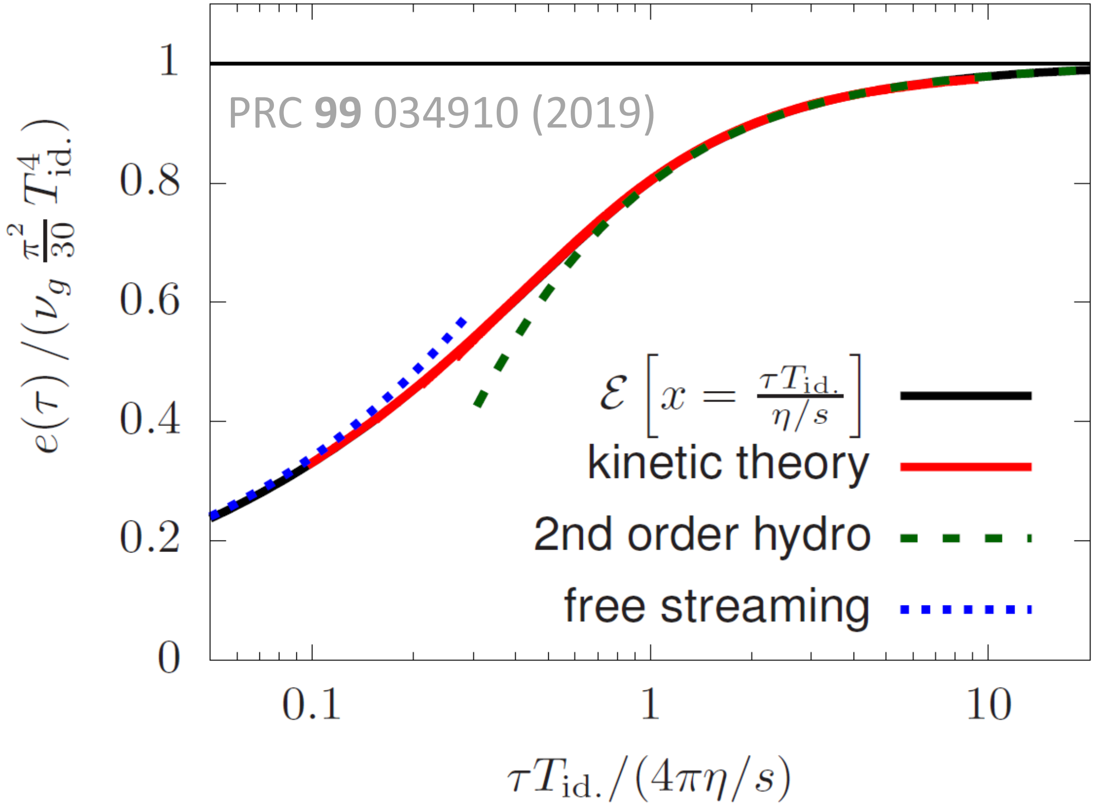
PRC **103** 054904 (2021)



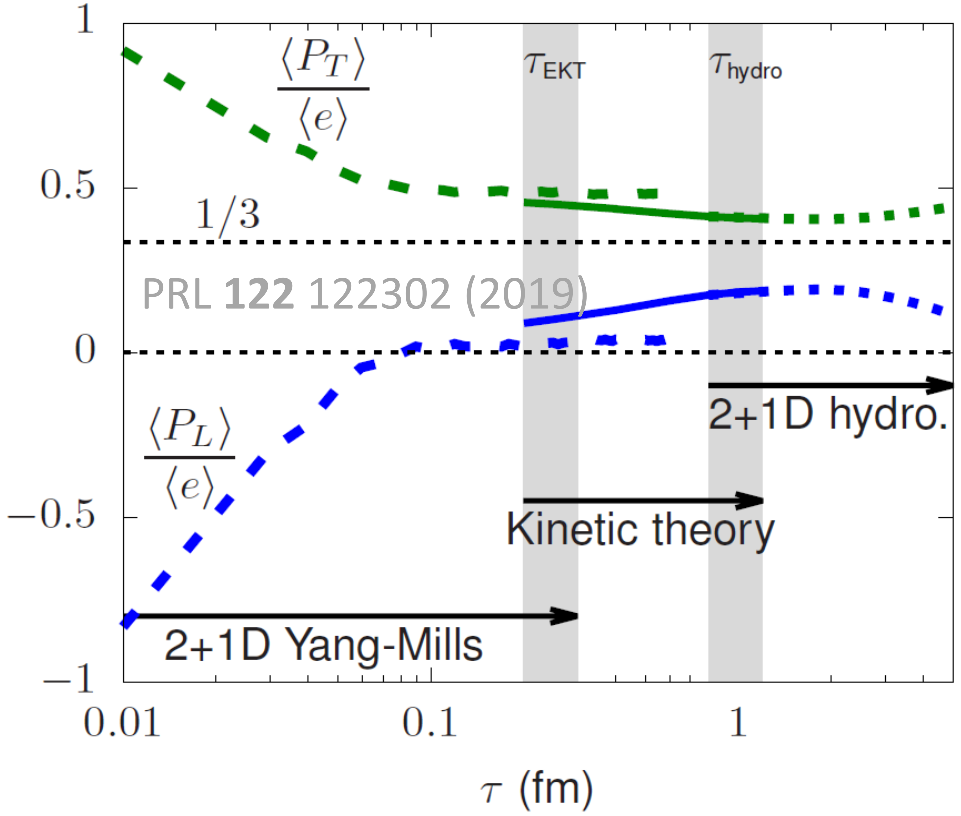
- Constraint on viscosities using RHIC and LHC data (at $\mu = 0$)
- $\mathcal{O}(1M)$ core-hours

Hydrodynamization of the Quark Gluon Plasma (QGP)

- Attractors are used to explore the approach towards hydrodynamical evolution of $T^{\mu\nu}$
  - Attractors: e.g. Arnold-Moore-Yaffe Effective Kinetic Theory of QCD and hydrodynamization (in 2+1D)



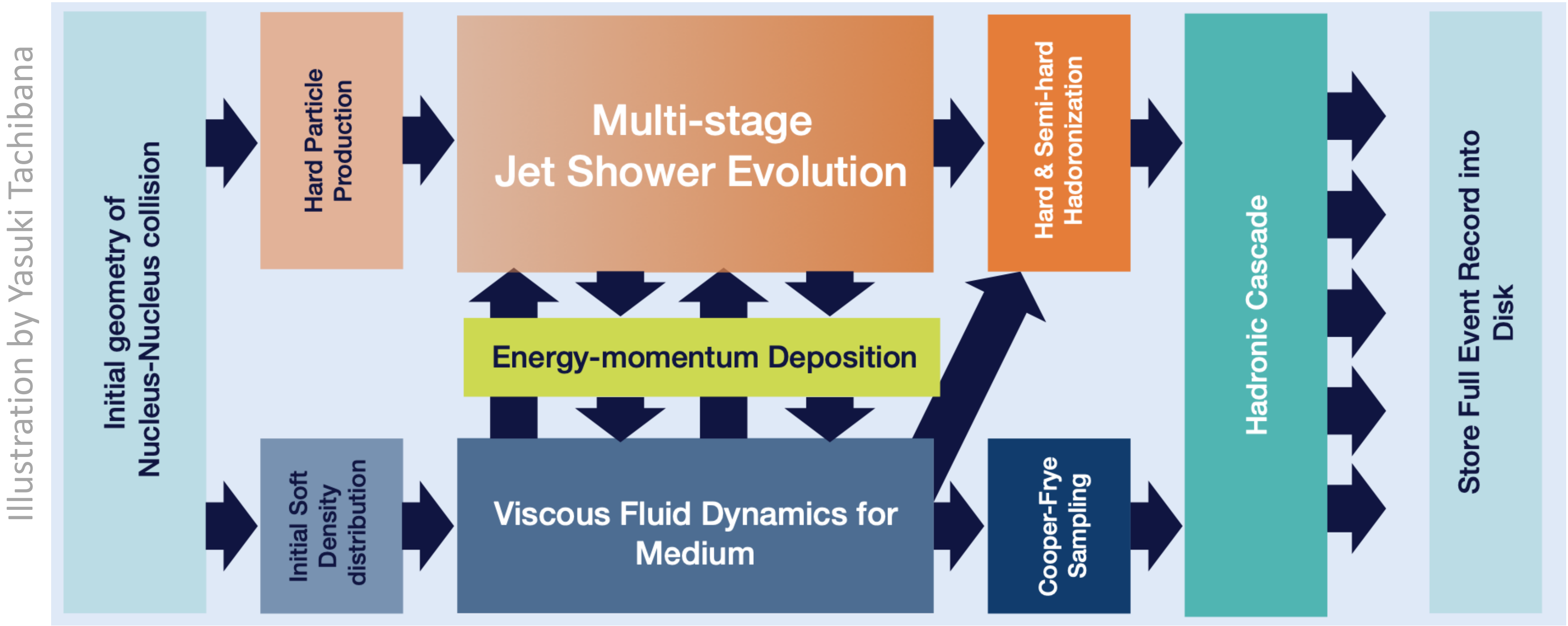
PRC **99** 034910 (2019)



PRL **122** 122302 (2019)

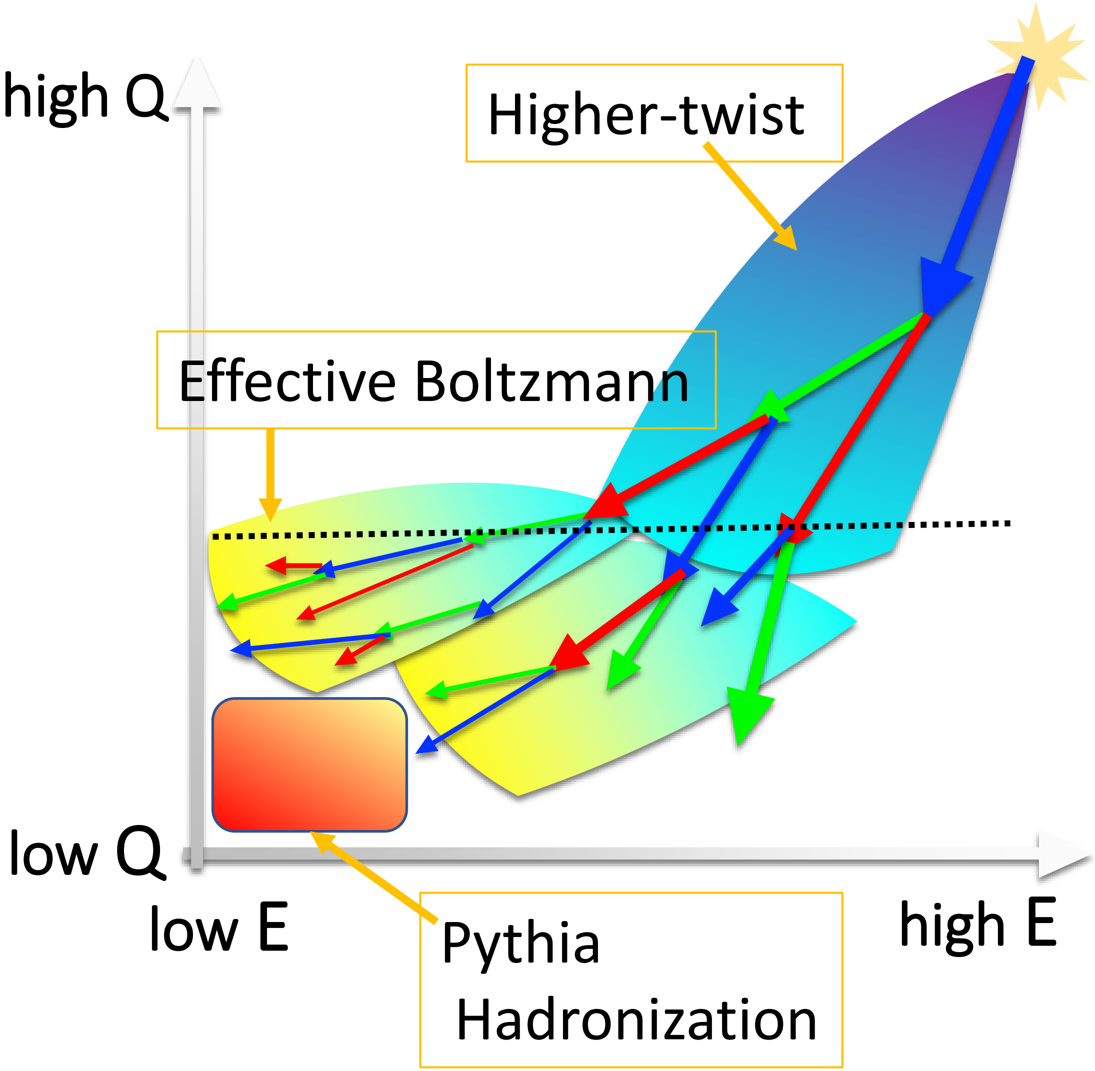
- Open question: how does quark flavor hydrodynamize?
  - Comparing FAIR vs RHIC/LHC allows to separate initial state quarks from dynamical quark-flavor production.
  - This goes in hand with exploring $P(T,\mu), \frac{\zeta}{s}(T,\mu), \frac{\eta}{s}(T,\mu)$ that will be done at facilities such as FAIR at GSI.

Simulating soft the nuclear medium



- JETSCAPE Collaboration provides a software framework describes:
  - Jets Monte Carlo event generators inside the nuclear medium allow more direct access to fundamental QCD degrees of freedom: the quarks and gluons ⇒ jets are sensitive to dynamical flavor hydrodynamization
  - Bayesian tools present within this framework can quantitatively study flavor hydrodynamization mechanisms.

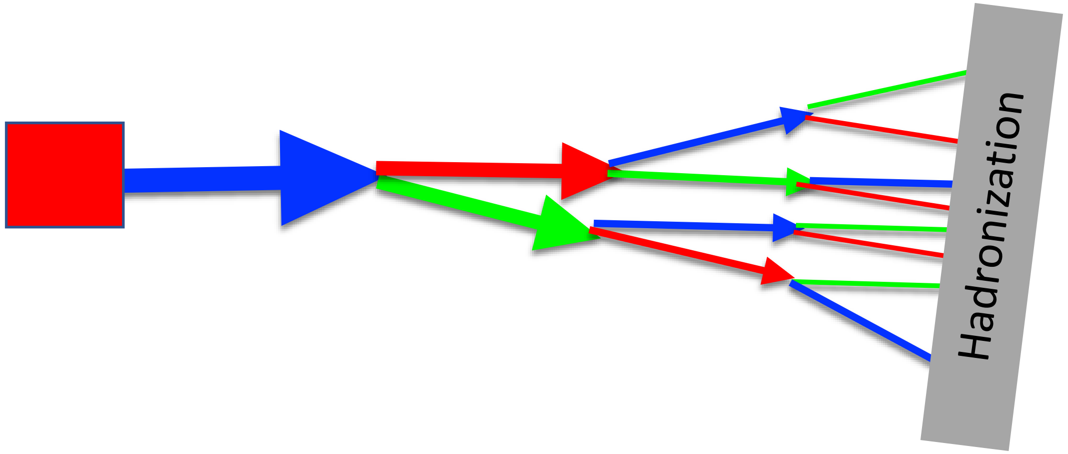
Multi-stage parton evolution in JETSCAPE



- High→Lower Q, High E: radiation is modified through single-scattering effects (using Higher-Twist approach)
- Low Q, High→Lower E: Multiple-scattering induced radiation (in effective or resummed Boltzmann Transport) plays a key role.
- Low Q, Low E: Hadronization physics important (non-perturbative)

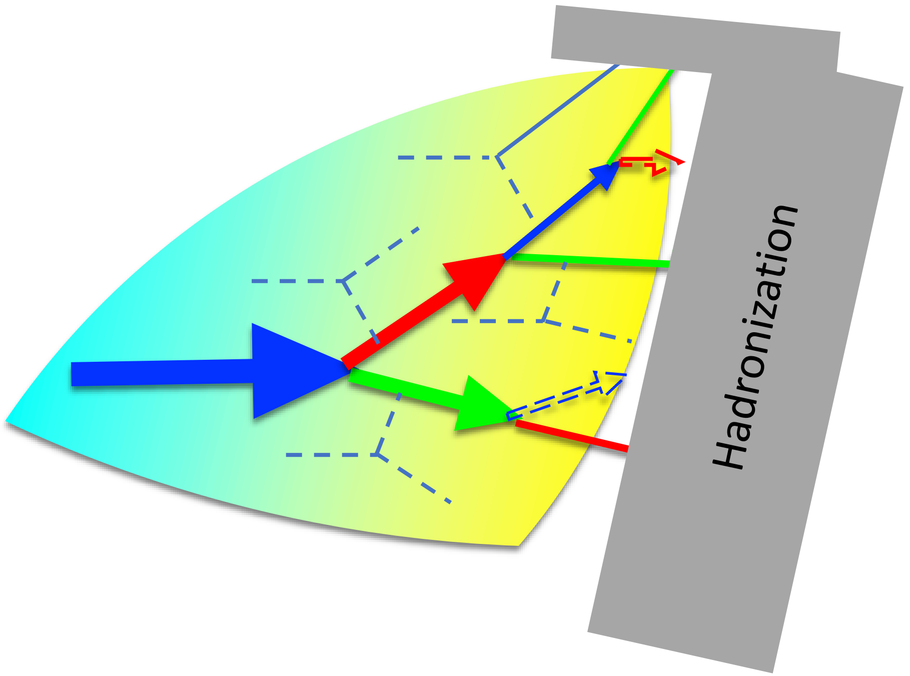
There is a sharp transition (at a constant Q^2) from Higher-Twist to effective Boltzmann ⇒ this approach needs to be revised at intermediate Q^2 by incorporating some multiple-scattering effects.

Monte Carlo jet shower simulation in vacuum



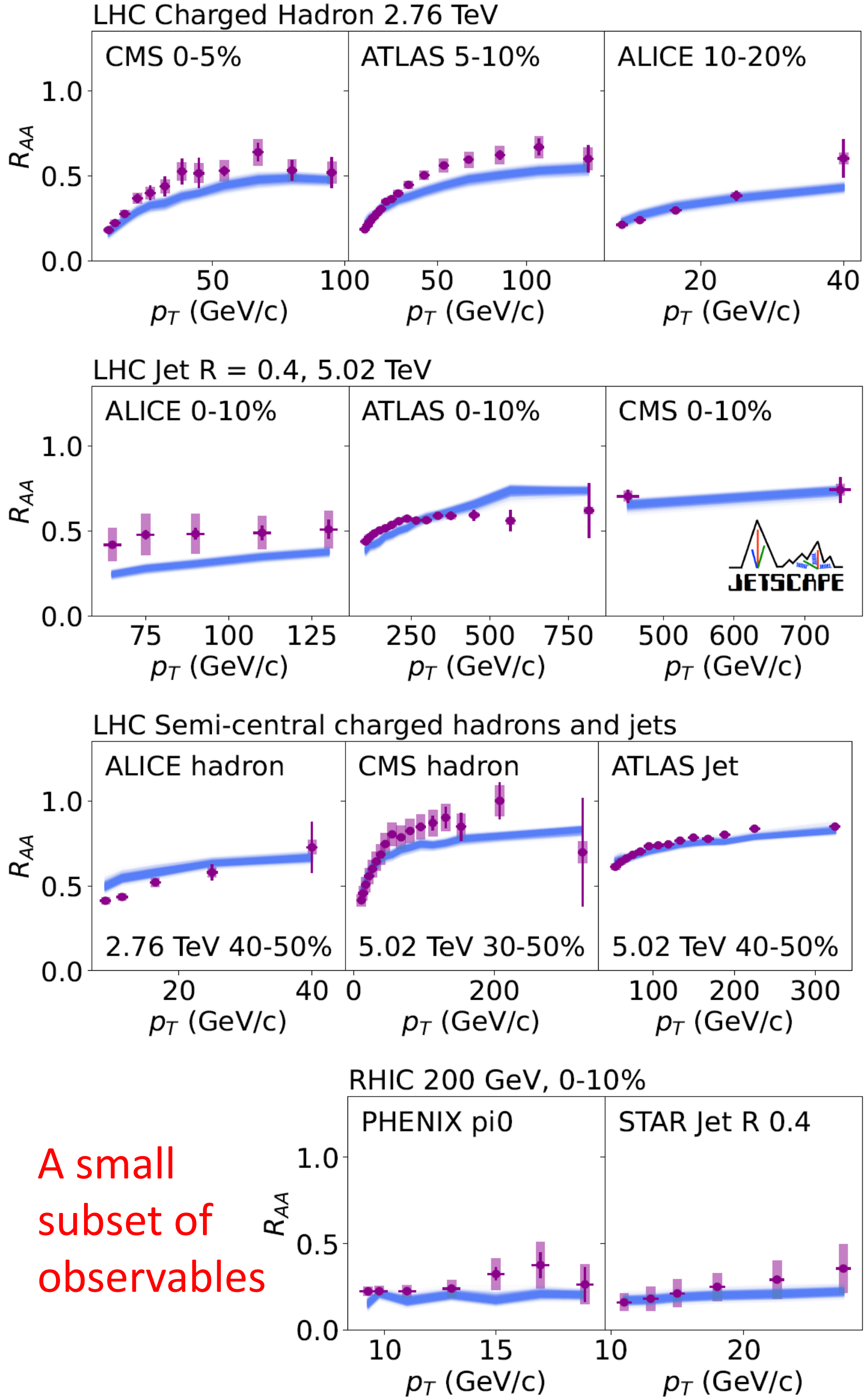
- Monte Carlo simulations (e.g., Pythia) develop a shower at the quark/gluon level in vacuum by adding multiple splits.
- In vacuum, the particles in a jet after hadronization occupy a narrow cone, as there is no effective Boltzmann transport.

Modified splitting inside the QGP

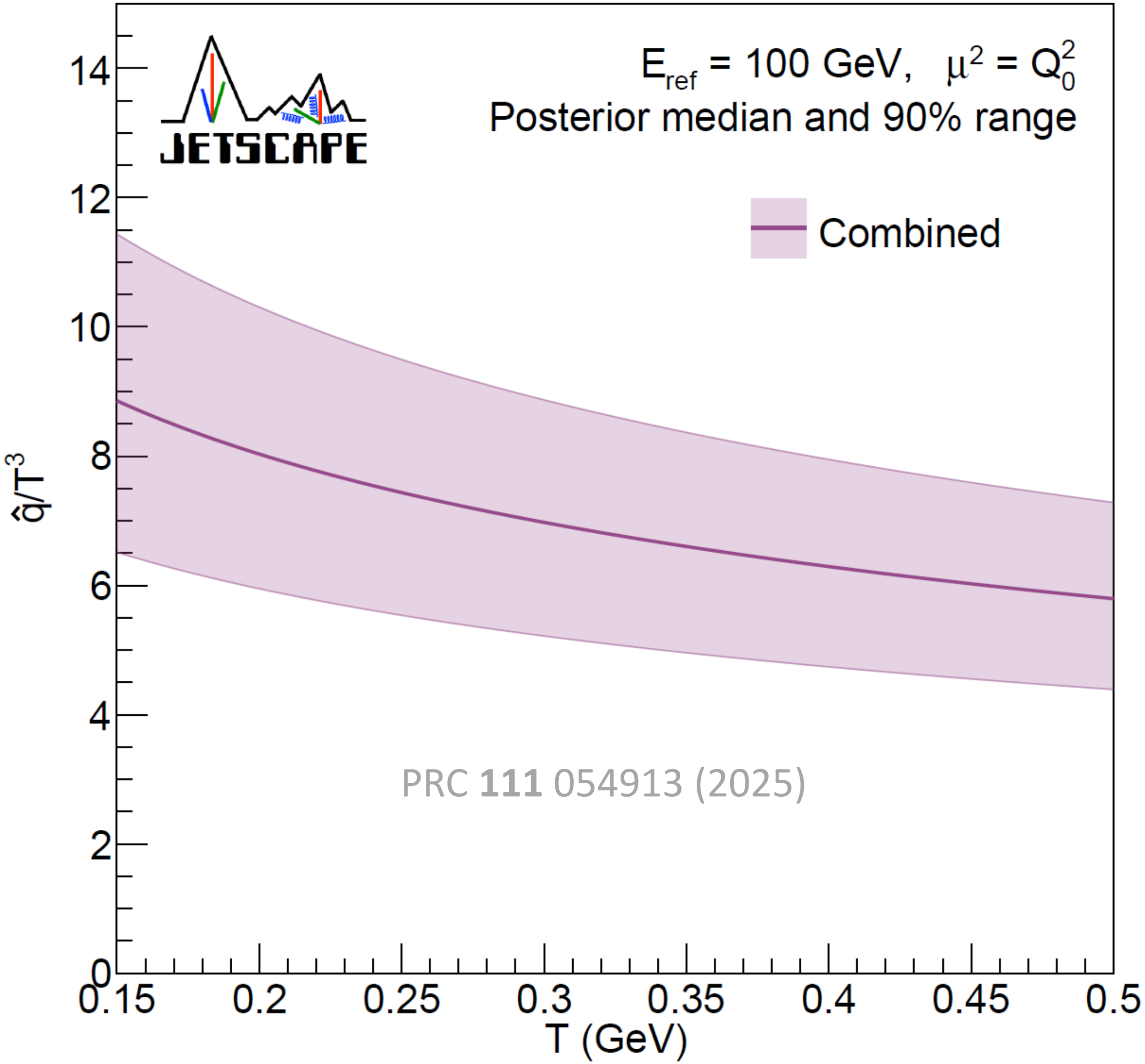


- In the nuclear medium, the particles in a jet after hadronization occupy a wider cone.
- The transport coefficient $\hat{q}$ is used to measure (transverse) momentum broadening in the QGP.
    - This transport coefficients is similar to the spatial Brownian diffusion coefficient.

JETS in the QGP



A small subset of observables



PRC **111** 054913 (2025)

- $\mathcal{O}(10M)$ core-hours used in this Bayesian analysis

Numerical simulations of heavy-ion collisions

- Separation of soft and hard physics:
    - Use simulations of the QGP fluid without jets to constrain some of its properties (e.g. viscosity) using Bayesian analysis.
    - Using the “best fit” of fluid parameters, i.e. Maximum A Posteriori likelihood, Bayesian analysis of jets have been used to constrain $\hat{q}$.

Numerical simulations of heavy-ion collisions

- Separation of soft and hard physics:
  - Use simulations of the QGP fluid without jets to constrain some of its properties (e.g. viscosity) using Bayesian analysis.
  - Using the “best fit” of fluid parameters, i.e. Maximum A Posteriori likelihood, Bayesian analysis of jets have been used to constrain $\hat{q}$.
- The next-generation of Bayesian analysis will simultaneously constrain jets and hydrodynamics, including jet energy deposition in the QGP. Such simulations would need $\mathcal{O}(1G)$ core-hours using current, mostly non-accelerated CPU-based approaches.

Numerical simulations of heavy-ion collisions

- Separation of soft and hard physics:
  - Use simulations of the QGP fluid without jets to constrain some of its properties (e.g. viscosity) using Bayesian analysis.
  - Using the "best fit" of fluid parameters, i.e. Maximum A Posteriori likelihood, Bayesian analysis of jets have been used to constrain $\hat{q}$.
- The next-generation of Bayesian analysis will simultaneously constrain jets and hydrodynamics, including jet energy deposition in the QGP. Such simulations would need $\mathcal{O}(1G)$ core-hours using current, mostly non-accelerated CPU-based approaches.
- Conventional approaches are to be revised and use more accelerated computing
  - Accelerators rely on Single Instruction Multiple Data (SIMD) paradigm to be used in simulation software

Numerical simulations of heavy-ion collisions

- Separation of soft and hard physics:
  - Use simulations of the QGP fluid without jets to constrain some of its properties (e.g. viscosity) using Bayesian analysis.
  - Using the "best fit" of fluid parameters, i.e. Maximum A Posteriori likelihood, Bayesian analysis of jets have been used to constrain $\hat{q}$.
- The next-generation of Bayesian analysis will simultaneously constrain jets and hydrodynamics, including jet energy deposition in the QGP. Such simulations would need $\mathcal{O}(1G)$ core-hours using current, mostly non-accelerated CPU-based approaches.
- Conventional approaches are to be revised and use more accelerated computing
  - Accelerators rely on Single Instruction Multiple Data (SIMD) paradigm to be used in simulation software

CPUs

  - Advanced Vector Instructions (AVX512): upto 32X double precision (64-bit), or upto 128X on half-precision (16-bit), computations on 1 CPU core
  - Advanced Matrix Instructions (4th gen Intel Xeons and beyond): upto 1024X speedup on matrix-matrix multiplications at 16-bit **only.** AMX is typically used to accelerate **AI** computations on CPUs, but can be *repurposed* when running physics simulations

Numerical simulations of heavy-ion collisions

- Separation of soft and hard physics:
  - Use simulations of the QGP fluid without jets to constrain some of its properties (e.g. viscosity) using Bayesian analysis.
  - Using the "best fit" of fluid parameters, i.e. Maximum A Posteriori likelihood, Bayesian analysis of jets have been used to constrain $\hat{q}$.
- The next-generation of Bayesian analysis will simultaneously constrain jets and hydrodynamics, including jet energy deposition in the QGP. Such simulations would need $\mathcal{O}(1G)$ core-hours using current, mostly non-accelerated CPU-based approaches.
- Conventional approaches are to be revised and use more accelerated computing
  - Accelerators rely on Single Instruction Multiple Data (SIMD) paradigm to be used in simulation software

CPUs

  - Advanced Vector Instructions (AVX512): upto 32X double precision (64-bit), or upto 128X on half-precision (16-bit), computations on 1 CPU core
  - Advanced Matrix Instructions (4th gen Intel Xeons and beyond): upto 1024X speedup on matrix-matrix multiplications at 16-bit **only.** AMX is typically used to accelerate **AI** computations on CPUs, but can be *repurposed* when running physics simulations
  - GPUs can give a larger acceleration, but are not necessarily the most practical, or efficient, for a given workload. GPUs need more power ⇒ fewer GPUs per datacenter rack

Numerical simulations of heavy-ion collisions

- Separation of soft and hard physics:
  - Use simulations of the QGP fluid without jets to constrain some of its properties (e.g. viscosity) using Bayesian analysis.
  - Using the “best fit” of fluid parameters, i.e. Maximum A Posteriori likelihood, Bayesian analysis of jets have been used to constrain $\hat{q}$.
- The next-generation of Bayesian analysis will simultaneously constrain jets and hydrodynamics, including jet energy deposition in the QGP. Such simulations would need $\mathcal{O}(1G)$ core-hours using current, mostly non-accelerated CPU-based approaches.
- Conventional approaches are to be revised and use more accelerated computing
  - Accelerators rely on Single Instruction Multiple Data (SIMD) paradigm to be used in simulation software

CPUs

    - Advanced Vector Instructions (AVX512): upto 32X double precision (64-bit), or upto 128X on half-precision (16-bit), computations on 1 CPU core
    - Advanced Matrix Instructions (4th gen Intel Xeons and beyond): upto 1024X speedup on matrix-matrix multiplications at 16-bit **only.** AMX is typically used to accelerate **AI** computations on CPUs, but can be *repurposed* when running physics simulations
    - GPUs can give a larger acceleration, but are not necessarily the most practical, or efficient, for a given workload. GPUs need more power ⇒ fewer GPUs per datacenter rack
- The next generation of large-scale simulations should better leverage SIMD-accelerated computing, at mixed/reduced precision – be it on CPUs, GPUs, AI accelerators – to keep the calculation time in check

Accelerated computing in Nuclear Physics

- New software writing paradigm: pack calculations in vectors/matrices to ease dynamical porting to the available accelerator hardware (be it AVX on CPUs, GPUs, or AI accelerators)
  - vendor agnostic packages such as Kokkos, OpenMP/ACC and so on can help, but they are not a silver bullet.
  - To ensure maximal SIMD acceleration scaling, good cache residency is crucial.

Accelerated computing in Nuclear Physics

- New software writing paradigm: pack calculations in vectors/matrices to ease dynamical porting to the available accelerator hardware (be it AVX on CPUs, GPUs, or AI accelerators)
  - vendor agnostic packages such as Kokkos, OpenMP/ACC and so on can help, but they are not a silver bullet.
  - To ensure maximal SIMD acceleration scaling, good cache residency is crucial.

- Beyond this, new communication protocols, such as Compute eXpress Link (CXL), allow software to decide on the combination of hardware to be used in a calculation.
  - Additional computational physics research is required to optimally leverage this technology, that allows more memory and more compute to be added, scaling beyond what is available within a node

Accelerated computing in Nuclear Physics

- New software writing paradigm: pack calculations in vectors/matrices to ease dynamical porting to the available accelerator hardware (be it AVX on CPUs, GPUs, or AI accelerators)
  - vendor agnostic packages such as Kokkos, OpenMP/ACC and so on can help, but they are not a silver bullet.
  - To ensure maximal SIMD acceleration scaling, good cache residency is crucial.

- Beyond this, new communication protocols, such as Compute eXpress Link (CXL), allow software to decide on the combination of hardware to be used in a calculation.
  - Additional computational physics research is required to optimally leverage this technology, that allows more memory and more compute to be added, scaling beyond what is available within a node
  - New computing hardware: mixes CPU, GPU, AI accelerators in a single socket (e.g. AMD MI300A, Intel Jaguar Shores,…)

⇒ Active research is needed to decide how best to combine different hardware via CXL

Accelerated computing in Nuclear Physics

- New software writing paradigm: pack calculations in vectors/matrices to ease dynamical porting to the available accelerator hardware (be it AVX on CPUs, GPUs, or AI accelerators)
  - vendor agnostic packages such as Kokkos, OpenMP/ACC and so on can help, but they are not a silver bullet.
  - To ensure maximal SIMD acceleration scaling, good cache residency is crucial.

- Beyond this, new communication protocols, such as Compute eXpress Link (CXL), allow software to decide on the combination of hardware to be used in a calculation.
  - Additional computational physics research is required to optimally leverage this technology, that allows more memory and more compute to be added, scaling beyond what is available within a node
  - New computing hardware: mixes CPU, GPU, AI accelerators in a single socket (e.g. AMD MI300A, Intel Jaguar Shores,…)

⇒ Active research is needed to decide how best to combine different hardware via CXL

- HPC endeavors currently requiring more than $\mathcal{O}(1G)$ core-hours, need new computer science languages to be co-devised to efficiently leverage new computing “devices” (*e.g. via CXL*), thus lessening the burden on simulation developers.
  - A combination of physicists, mathematicians, and computer scientists, as well as industry partners should be established to achieve this goal (see arxiv:2501.00905).

Accelerated computing in Nuclear Physics

- Devising sophisticated HPC code also implies software stewardship.

- A survey of many fields in nuclear physics (arxiv:2501.00905) has identified the need for dedicated software stewards, to ensure future generations can effectively use and upgrade the software infrastructure in various nuclear physics fields

- A new category of physicist, i.e. computational physicist or data physicist (arxiv:2501.00905), is suited to "shepherd" stewardship demands, while ensuring a software that leverages newest computing paradigms, in both theoretical and experimental physics

- For a timely realization of next-generation of large-scale computing endeavors, such as large-scale Bayesian analysis, physicists should be actively involved in computing research.
- Research in accelerated computing, such as development of new computing languages, is conducive to establish new partnerships with industry, which uses accelerated computing in AI.

Conclusion and Outlook

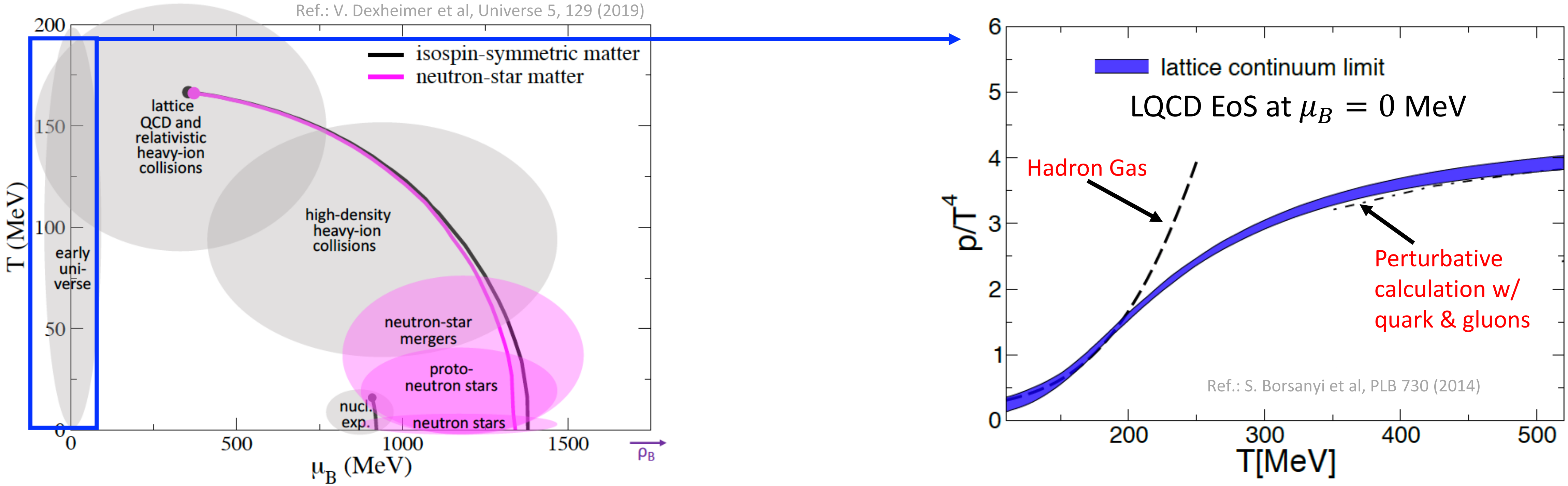
- Using Bayesian analysis constraints on QGP viscosities and jet momentum diffusion was obtained.
- Theory development is needed in:
  - Flavor hydrodynamization
  - Multi-scale modelling of jets: a better “interpolation” between high- and low-Q^2 evolution
- The next generation of Bayesian analysis will simultaneously constrain jets and the QGP fluid, requiring $\geq$100X more computing power.
- To reach that goal, accelerated computing, especially at mixed and reduced precision, is needed (arxiv:2501.00905)
  - SIMD computing paradigm (on CPUs, GPUs and AI accelerators) allows to speed-up execution of code, keeping the simulation time in check
  - Computing research (including languages) leveraging any/all available hardware, together with new communication protocols CXL that can reconfigure the hardware for particular needs, should enabled in larger-scale scientific discoveries using HPC
- Great possibilities for physicists, mathematicians, computer scientist to work together and tackle numerically challenging problems, with great outcome (arxiv:2501.00905)
  - A computational physicist or a data physicist is suited for such a task: being at the forefront of HPC research software development and stewardship
  - New collaboration with HPC industry partners speeds-up development and adoption of accelerated computing, used in AI.

Backup

Nuclear equation of state in thermal equilibrium

- $\uparrow \sqrt{s_{NN}} \Rightarrow$ more gluons $\Rightarrow n_q \sim n_{\bar{q}} \Rightarrow \mu_B \approx 0$
- $\downarrow \sqrt{s_{NN}} \Rightarrow$ more valance quarks $\Rightarrow \mu_B > 0$

Lattice QCD (L-QCD) equation of state (EoS)



Ref.: V. Dexheimer et al, Universe 5, 129 (2019)

Ref.: S. Borsanyi et al, PLB 730 (2014)

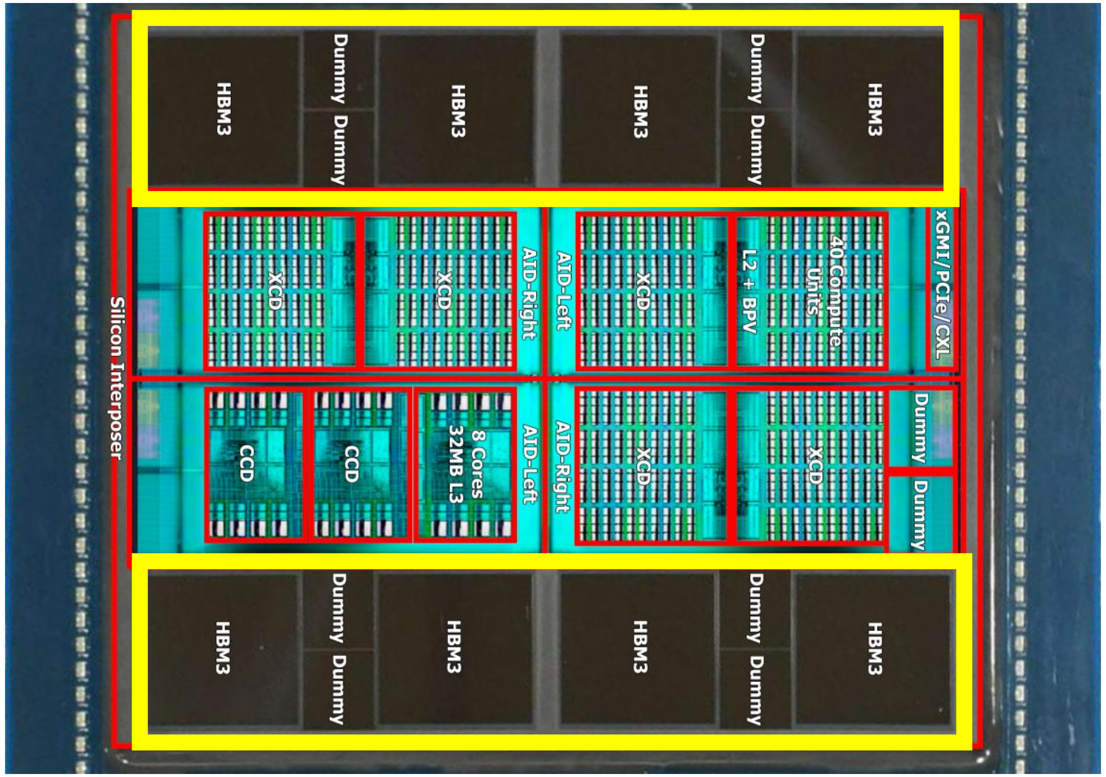
- What is $P(T, \mu_B)$?
  - Lattice QCD with Taylor expansion around $\mu_B = 0$
  - Perturbative QCD calculations for $\mu_B > 0$
  - Hadronic calculations for $\mu_B > 0$ are lacking

Next step: A simultaneous Bayesian calibration of jets and fluid dynamics

- CPU have accelerators (AVX, AMX), so jet simulations need to be optimized to exploit it.
- For hydrodynamics, CPU accelerators help, but realistically GPUs are needed.

⇒ Ideally a GPU-CPU hybrid is required

- AMD MI300A a CPU-GPU hybrid architecture



MI300A from AMD

- One 24-core CPU tile (highlighted in yellow)
- 3 GPU tiles (surrounding the CPU tile)
- To and bottom rows of tiles: 128GB of High-Bandwidth Memory (HBM)

Intel is working on something similar in its upcoming Jaguar Shores processor.

https://semianalysis.com/2023/06/12/amd-mi300-taming-the-hype-ai-performance/

An irreducible tensor decomposition of hydrodynamics

- In high-energy collisions (w/ negligible μ_B), what is flowing?... That can only be energy density ϵ

$$T^{\mu\nu}u_\nu = \epsilon u^\mu$$

Landau's flow definition

$u^\mu = \left(\gamma, \gamma\vec{\beta}\right)$ where

$\gamma = \left(1-\beta^2\right)^{-1/2}$ and $\vec{\beta} = \vec{v}/c$. Using natural units from how on $\Rightarrow c = 1$.

- Non-dissipative $T_0^{\mu\nu}$ can only take the form:

$$T_0^{\mu\nu} = \epsilon u^\mu u^\nu - P(\epsilon)\Delta^{\mu\nu} = \epsilon u^\mu u^\nu - P(\epsilon)(g^{\mu\nu} - u^\mu u^\nu)$$

- Including dissipation gives rise to dissipative corrections $\delta T^{\mu\nu}$ to $T_0^{\mu\nu}$, namely Π and $\pi^{\mu\nu}$

$$T^{\mu\nu} = T_0^{\mu\nu} + \delta T^{\mu\nu} = T_0^{\mu\nu} - \Pi\Delta^{\mu\nu} + \pi^{\mu\nu} = \epsilon u^\mu u^\nu - (P + \Pi)\Delta^{\mu\nu} + \pi^{\mu\nu}$$

where the viscous pressure are decomposed in terms of irreducible tensors, namely

radial deformations

$$\Pi = -\tfrac{1}{3}\Delta^{\mu\nu}T_{\mu\nu} - P(\epsilon)$$

angular deformations

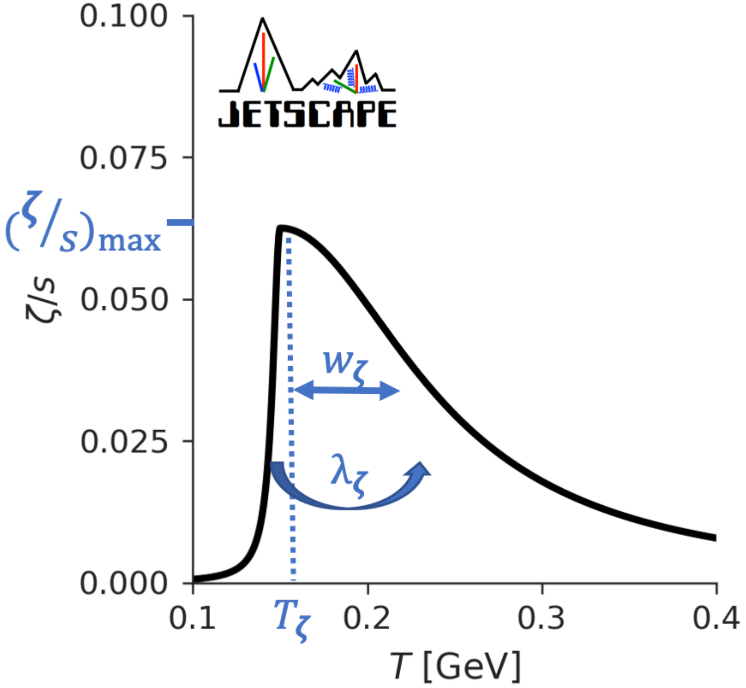
$$\pi^{\mu\nu} = T^{\langle\mu\nu\rangle} = \Delta^{\mu\nu}_{\alpha\beta}\,T^{\alpha\beta} = \left[\frac{1}{2}\left(\Delta^\mu_\alpha\Delta^\nu_\beta + \Delta^\mu_\beta\Delta^\nu_\alpha\right) - \frac{1}{3}\Delta^{\mu\nu}\Delta_{\alpha\beta}\right]T^{\alpha\beta}$$

w/ $\pi^\mu_\mu = 0$ and $u_\mu\pi^{\mu\nu} = 0$

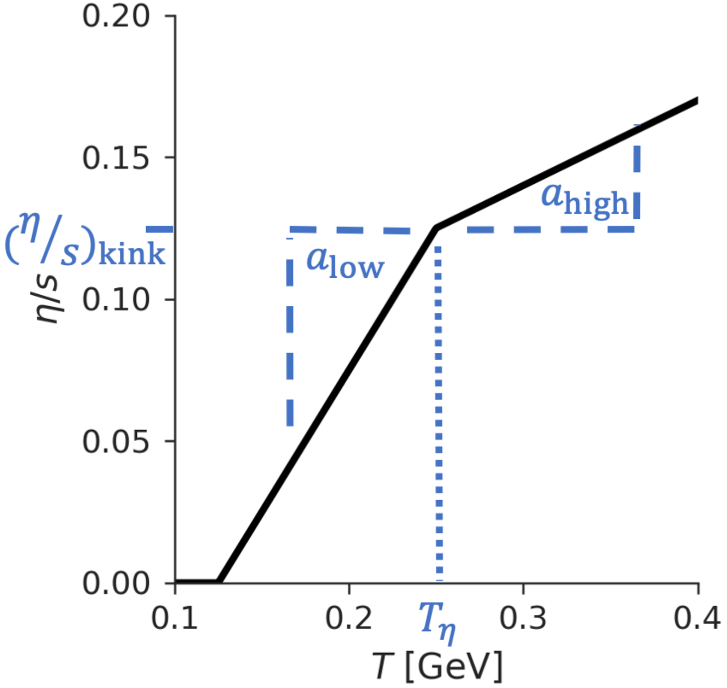
Modelling specific bulk (ζ/s) and shear (η/s) viscosities

- Bulk and shear viscosities were parametrized using 4-parameter functions

PRC **103** 054904 (2021)



PRC **103** 054904 (2021)



$$\frac{\zeta}{s}(T) = \frac{\left(\frac{\zeta}{s}\right)_{\max}\Lambda^2}{\Lambda^2 + \left(T - T_\zeta\right)^2}$$

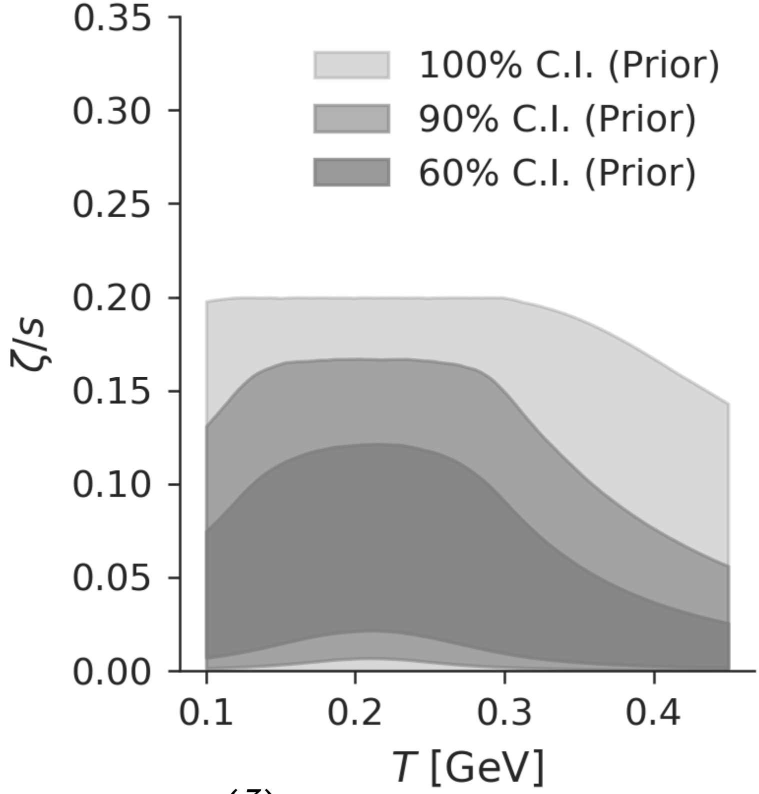
$$\Lambda = w_\zeta\left[1 + \lambda_\zeta\left(T - T_\zeta\right)\right]$$

$$\frac{\eta}{s}(T) = a_{\text{low}}\left(T - T_\eta\right)\Theta\left(T_\eta - T\right) + \left(\frac{\eta}{s}\right)_{\text{kink}} + a_{\text{high}}\left(T - T_\eta\right)\Theta\left(\mathrm{T} - \mathrm{T}_\eta\right)$$

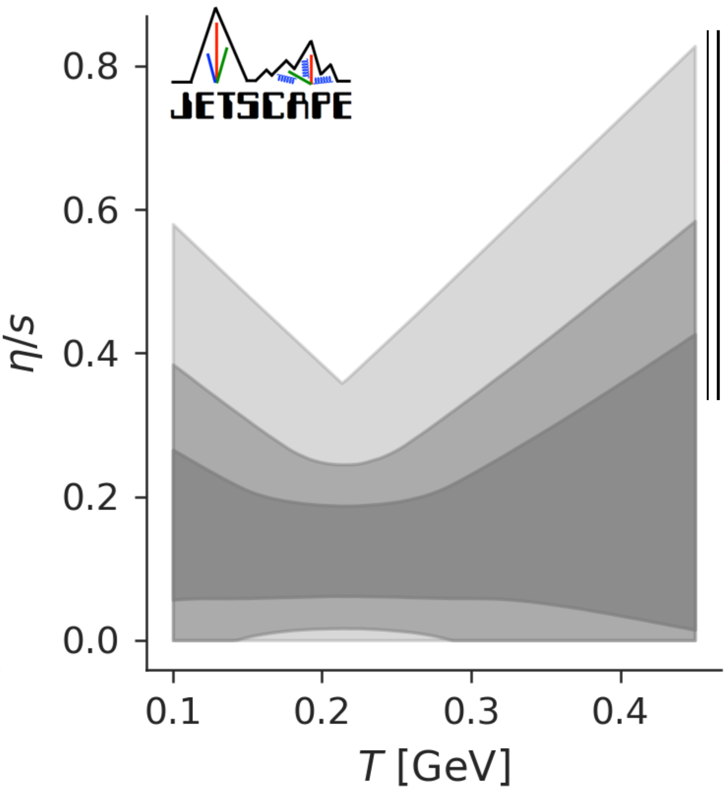
Modelling specific bulk (ζ/s) and shear (η/s) viscosities

- Bulk and shear viscosities were parametrized using 4-parameter functions

PRC **103** 054904 (2021)



PRC **103** 054904 (2021)



PRC **103** 054904 (2021)

temperature of (η/s) kink	T_η	[0.13, 0.3] GeV
(η/s) at kink	$(\eta/s)_{\mathrm{kink}}$	[0.01, 0.2]
low temp. slope of (η/s)	a_{low}	[–2, 1] GeV^{-1}
high temp. slope of (η/s)	a_{high}	[–1, 2] GeV^{-1}
shear relaxation time factor	b_π	[2, 8]
maximum of (ζ/s)	$(\zeta/s)_{\mathrm{max}}$	[0.01, 0.25]
temperature of (ζ/s) peak	T_ζ	[0.12, 0.3] GeV
width of (ζ/s) peak	w_ζ	[0.025, 0.15] GeV
asymmetry of (ζ/s) peak	λ_ζ	[–0.8, 0.8]

$$\frac{\zeta}{s}(T) = \frac{\left(\frac{\zeta}{s}\right)_{\mathrm{max}} \Lambda^2}{\Lambda^2 + \left(T - T_\zeta\right)^2}$$

$$\Lambda = w_\zeta\left[1 + \lambda_\zeta\left(T - T_\zeta\right)\right]$$

$$\frac{\eta}{s}(T) = a_{\mathrm{low}}\left(T - T_\eta\right)\Theta\left(T_\eta - T\right) + \left(\frac{\eta}{s}\right)_{\mathrm{kink}} + a_{\mathrm{high}}\left(T - T_\eta\right)\Theta\left(\mathrm{T} - \mathrm{T}_\eta\right)$$