Research Data in Lattice Quantum Field Theory

Bartosz Kostrzewa Carsten Urbach

Research in Lattice Quantum Field Theory (LQFT) is performed by international collaborations in both overlapping and disjoint research projects studying a wide range of non-perturbative physical problems. The generated data, which can roughly be classified into three tiers, span the whole spectrum of storage, metadata and lifetime requirements.

LQFT simulations are some of the most expensive in computational science and there are efforts underway to make some of the research data accessible in a FAIR way. The challenges to be overcome to get there are substantial.

Lattice 2022, Bonn

Lattice gauge ensembles and data management

https://arxiv.org/abs/2212.10138

Lattice 2022, Bonn

The International Lattice Data Grid towards FAIR Data

https://arxiv.org/abs/2212.08392

PUNCH4NFDI

Particles, Universe, NuClei and Hadrons for the NFDI

https://www.punch4nfdi.de

Hamiltonian Monte Carlo

tmLQCD software suite

https://github.com/etmc/tmLQCD

LEMON MPI I/O for LQCD

https://github.com/etmc/lemon

hadron LQCD analysis package for R

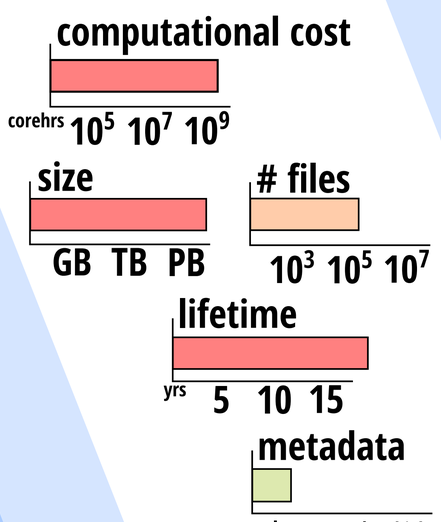
https://github.com/hiskp-lqcd/hadron

primary

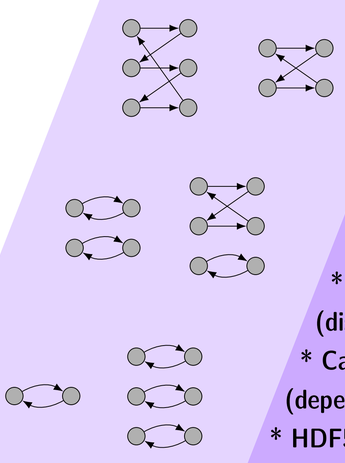
The starting point for any calculation in LQFT is a so-called ensemble of gauge configurations U, on which the Euclidean path integral of some observable O is evaluated using importance sampling.

$$\langle \hat{O} \rangle = \int \mathcal{D}U \ \det(D[U]) \ \hat{O} \ e^{-S[U]} = \sum_{i=1}^{N} O(U_i) + \mathcal{O}(1/\sqrt{N})$$

* Configurations U_i generated on supercomputers using the Hamiltonian Monte Carlo (HMC) algorithm.
* Tens of ensembles with thousands of configurations each required to evaluate observables.
* Total computational cost: billions of core hours, PB-level archival requirements for many years.
* Storage mostly via yearly data project applications at supercomputing sites or local infrastructure.
* LEMON MPI-I/O library for writing and reading using tens to thousands of MPI tasks.
* PUNCH4NFDI is working to revive the International Lattice Data Grid (ILDG) to provide infrastructure and middleware for annotation, archival, search and retrieval and to make configurations available to community long-term in a FAIR way.



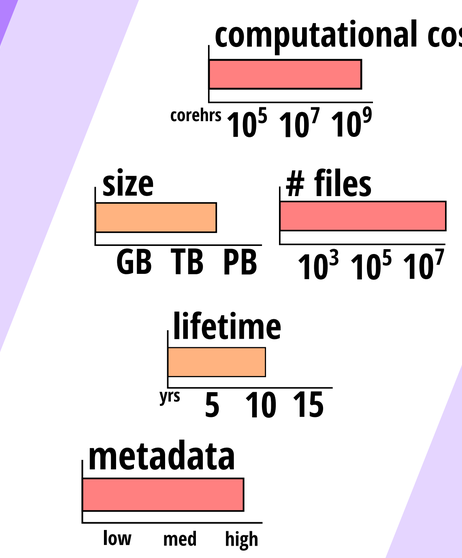
secondary



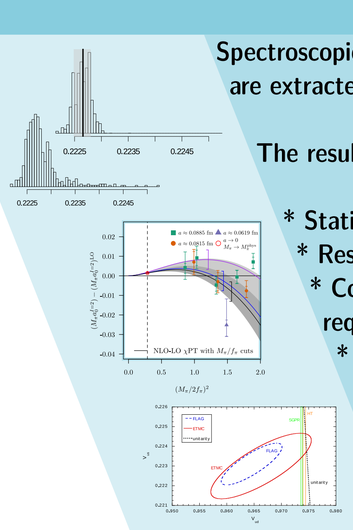
LQFT observables: correlation functions which probe the interaction between different particles, allowing their properties to be determined.

$$C(x,y,z) = \mathrm{tr}\left[\{D^{-1}_{\chi_d}(z \to x)\}^{\tilde{\dagger}} \gamma_5 \Gamma_c D^{-1}_{\chi_u}(x \leftarrow y) \Gamma_f D^{-1}_{\chi_d}(y \leftarrow z) \Gamma_i \gamma_5\right]$$

* D^{-1}: inverse of very large, very sparse matrix (expensive).
* C(x,y,z): tensor contraction of products of such matrices.
* Evaluated hundreds to thousands of times per configuration U.
* I/O excessive for D^{-1}: on-the-fly calculations or compression (distillation).
* Careful planning required to compute many different C at the same time (depending on various physical parameters).
* HDF5 used by many groups for hierarchical storage.
* Most datasets not FAIR, few storage standards.



tertiary

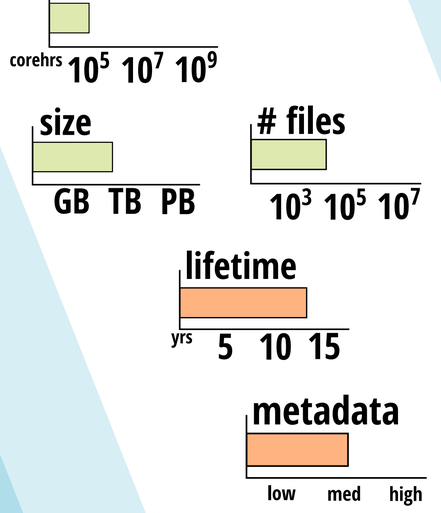


Spectroscopic information (masses, energies) and couplings (matrix elements) are extracted from correlation functions in various limits.

The results are subject to further analysis, which in itself produces research data:

* Statistical uncertainties are determined using resampling (jackknife/bootstrap).
* Results may have to be combined or inter-/extrapolated.
* Comparison to experiment or construction of dependent observables may require parametric modelling (fits).
* Taking into account systematic uncertainties may require model averaging.

Results are often published without the resulting artifacts or the corresponding analysis workflows and most groups have not adopted processes to preserve them long-term. These aspects make reproducibility difficult.



UNIVERSITÄT BONN

HPC/A-LAB

HISKP

Extended Twisted Mass Collaboration

TeRAbytes Workshop 17-18 Jan 2023 Universität Bonn