Status of Lattice Use Case via REANA:

Simran Singh
(University of Bielefeld)

with Olaf Kaczmarek , Ding-Ze Hu (work on Metadata) and help from Elena Sacchi -(Reana)

9th PUNCH4NFDI General Meeting

June 20th, 2024 Thursday 11.00

UNIVERSITÄT BIELEFELD
Faculty of Physics

Background image : DALL.E

Heavy-Quark Diffusion coefficient from Lattice simulations

- Luis Altenkort, Olaf Kaczmarek, Rasmus Larsen, Swagato Mukherjee, Peter Petreczky, Hai-Tao Shu, and Simon Stendebach (HotQCD Collaboration)
Phys. Rev. Lett. 130, 231902

Why is this a good candidate for a Lattice QCD use-case?

- Lattice QCD is currently our best hope to understand low-energy QCD - because the theory is strongly interacting.
- Particularly, understanding the hydrodynamic behaviour of quark gluon plasma (QGP) is important (even for experimental particle physicists).
- The heavy quark diffusion coefficient κ captures the momentum transfer to the heavy quark from the QGP background.
- Until now, due to computational challenges of simulating dynamical quarks, only non-dynamical quarks simulations were used to determine κ.
- For the first time, dynamical quarks have been used in such a calculation for κ

Simulation set-up & Computing resources

Gauge configurations generated from LQCD simulations of N_f = 2+1 at pion mass 320 MeV using SIMULATeQCD : a multi-GPU C++ code public & published : `https://latticeqcd.github.io/SIMULATeQCD/`

To be made available via ILDG

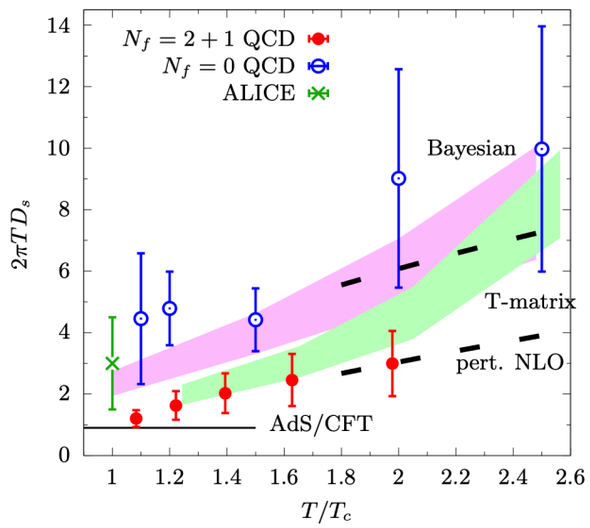
Computing resources used :

- **Bielefeld GPU cluster**
- **JUWELS at GCS@FZJ**
- **Marconi 100 at CINECA**

Raw Gauge configurations used to calculate Observable of interest : chromoelectric correlator (G_E) for various Temperatures

PUNCH output

Analysis Toolbox : Collection of Python tools developed at Bielefeld (Public): `https://github.com/LatticeQCD/AnalysisToolbox`



*Original full workflow and related data (expect gauge configuration which will eventually be made available via ILDG) can be found @ https://doi.org/10.4119/unibi/2979080

General outline of the workflow and file structure

✳ The main script called “HQDworkflow.sh” un-tars the input data calls multiple scripts in the following order.

1. **1_merge_data.sh :** Data from small text files merged into large binary files. Along with data, metadata is stored in a file containing information on number of files per stream, MCMC trajectory number, etc

2. **2_reduce_data.sh :** Merged data files loaded, extract an equally spaced MCMC time series, then plot the time history of the Polyakov loop at a large flow time.

3. **3_spline_interpolate.sh :** Interpolate the EE correlator in Euclidean time and in flow time, such that a common set of normalized flow times is available across all lattices and temperatures

4. **4_continuum_extr.sh :** Take the continuum limit of the EE correlator using a combined fit on each sample

5. **5_flowtime_extr.sh :** Take the flow-time-to-zero limit of the EE correlator using a combined fit on each sample

6. (!computation heavy!) Spectral function reconstruction

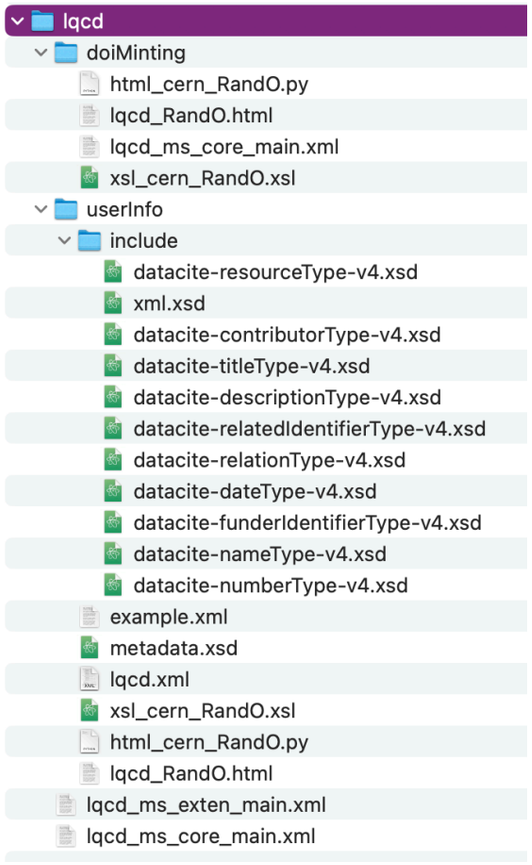
7. Comparison of results - final figure of paper

*Documentation and files suitable for REANA can be found at https://gitlab-p4n.aip.de/ssingh/latticeusecaseforscitrace

Work on Lattice UseCase by Ding-Ze Hu

Metadata Schema

- Designed metadata schema from DataCite and applied to lattice use-case. ✓
- Type A for DOI minting: Only contains the necessary information that helps us to find and identify the information. ✓
- Type B for users : contains information like size, format of the data, related information (software/article/...) and description of the data (methods/technical info/…) ✓
- All the XML metadata records are tested to ensure that they are valid against the schema from DataCite using xmllint tool ✓

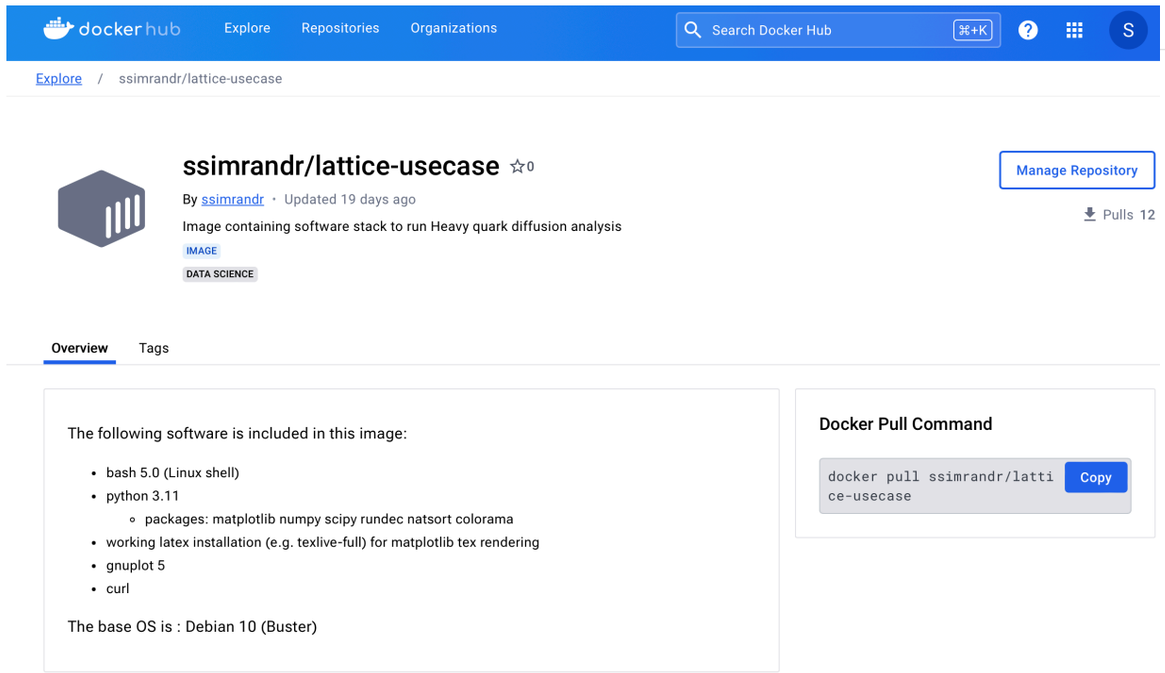


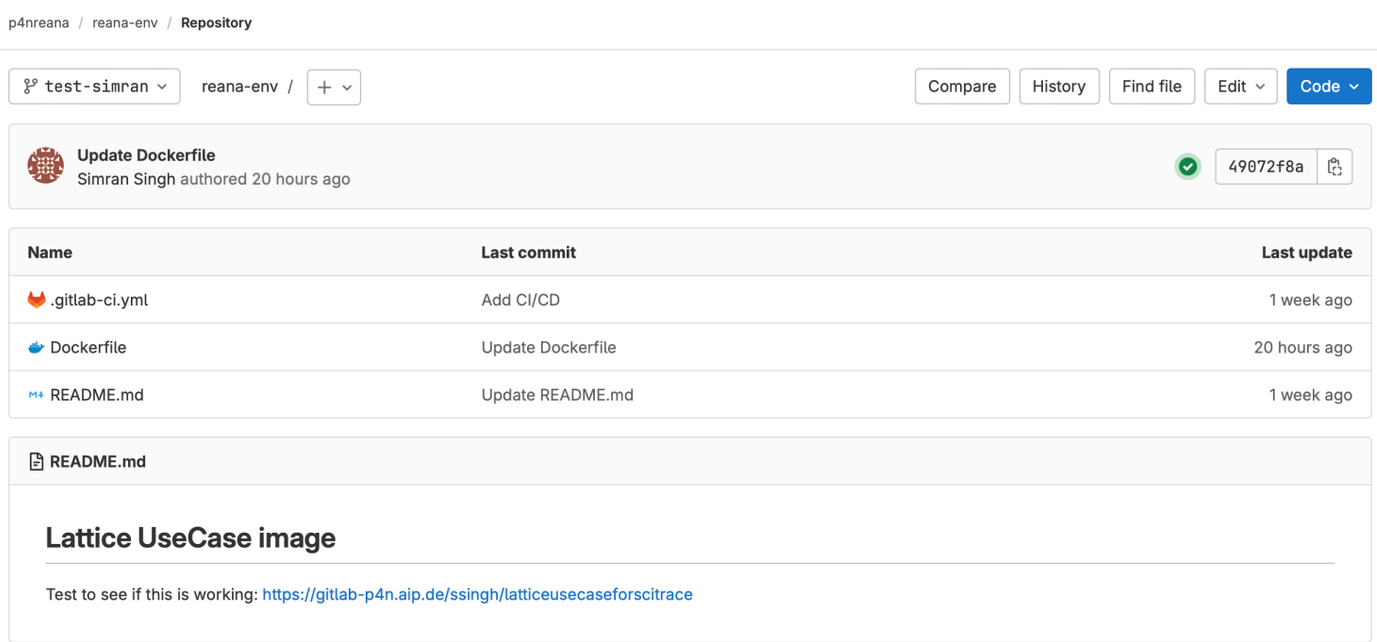
Docker Container run on Compute4Punch

- Created a Docker container specifically for the Lattice use case. ✓
- The container includes all necessary data and programs for running LQCD analysis and generating result charts. ✓
- Registered within the Compute4PUNCH platform, enabling the execution of the entire LQCD analysis workflow. ✓

Status on running the workflow on REANA : DockerEnvironment

Required Software and Environment via Docker Image at registry





- Publicly available to pull and build the use case ✓
- However, when used with REANA - compatibility issues ✗
- Elena managed to make it work using GitLab's CI/CD pipeline ✓

Status on running the workflow on REANA : WorkFlow YAML file

reana.yaml 488 B

```yaml
inputs:
  files:
    - download_data.sh
    - HQDworkflow.sh
  parameters:
    num_processes: 40
workflow:
  type: serial
  specification:
    steps:
      - environment: 'gitlab-p4n.aip.de:5005/p4nreana/reana-env:test-simran.49072f8a'
        commands:
        - chmod +x download_data.sh
        - chmod +x HQDworkflow.sh
        - ./download_data.sh
        - mkdir -p output
        - ./HQDworkflow.sh ${num_processes}
outputs:
  files:
    - output/*.pdf
  directories:
    - output

```

- Cannot store such large input files on gitlab repository - hence need to download input data and software

Status on running the workflow on REANA : Results of workflow

- First run resulted in failure due to OOM issues.
- One part of the workflow is extremely memory and resource intensive (spectral function reconstruction)

```
latticeusecaseforscitrace — -zsh — 131×35
Initial guess for fit params: [1, 1]
Bounds for fit params:  [[0.01, 20], [0.01, 5]]
./correlators_flow/spf_reconstruction/model_fitting/example_usage/spf_reconstruct.sh: line 15:  6230 Killed                  "$@"
Matplotlib created a temporary cache directory at /tmp/matplotlib-o93znyme because the default path (/.config/matplotlib) is not a
writable directory; it is highly recommended to set the MPLCONFIGDIR environment variable to a writable directory, in particular to
 speed up the import of Matplotlib and to better support multiprocessing.

2024/06/17 20:00:26
../spf_reconstruct.py --output_path /var/reana/users/86c2a219-bcca-49b5-85ea-b6f8c1c90278/workflows/f82f5854-bb98-426c-ad2c-a87aeaf
63db3/output_data/hisq_ms5_zeuthenFlow/EE//T195/spf/ --add_suffix 23-02-16_relflow --input_corr /var/reana/users/86c2a219-bcca-49b5
-85ea-b6f8c1c90278/workflows/f82f5854-bb98-426c-ad2c-a87aeaf63db3/output_data/hisq_ms5_zeuthenFlow/EE//T195/EE_flow_extr_relflow.np
y --min_tauT 0.24 --nproc 40 --T_in_GeV 0.195 --corr_from_combined_fit_nt 36 --Nf 3 --nsamples 1000 --model plaw --PhiUV_order LO -
-omega_prefactor 1 --min_scale eff --OmegaByT_IR 1 --OmegaByT_UV 6.2832

min_scale =  1.771
omega_prefactor =  1.000
scale > min_scale at OmegaByT= 9.083
Initial guess for fit params: [1, 1]
Bounds for fit params:  [[0.01, 20], [0.01, 5]]
./correlators_flow/spf_reconstruction/model_fitting/example_usage/spf_reconstruct.sh: line 15:  6298 Killed                  "$@"
Matplotlib created a temporary cache directory at /tmp/matplotlib-ksp5cj9i because the default path (/.config/matplotlib) is not a
writable directory; it is highly recommended to set the MPLCONFIGDIR environment variable to a writable directory, in particular to
 speed up the import of Matplotlib and to better support multiprocessing.
Matplotlib is building the font cache; this may take a moment.

2024/06/17 20:01:12
../spf_reconstruct.py --output_path /var/reana/users/86c2a219-bcca-49b5-85ea-b6f8c1c90278/workflows/f82f5854-bb98-426c-ad2c-a87aeaf
63db3/output_data/hisq_ms5_zeuthenFlow/EE//T195/spf/ --add_suffix 23-02-16_relflow --input_corr /var/reana/users/86c2a219-bcca-49b5
-85ea-b6f8c1c90278/workflows/f82f5854-bb98-426c-ad2c-a87aeaf63db3/output_data/hisq_ms5_zeuthenFlow/EE//T195/EE_flow_extr_relflow.np
y --min_tauT 0.24 --nproc 40 --T_in_GeV 0.195 --corr_from_combined_fit_nt 36 --Nf 3 --nsamples 1000 --model plaw --PhiUV_order NLO
--omega_prefactor opt --min_scale eff --OmegaByT_IR 1 --OmegaByT_UV 6.2832

min_scale =  1.771
omega_prefactor =  14.743
scale > min_scale at OmegaByT= 0.616
```

→ OOMKilled

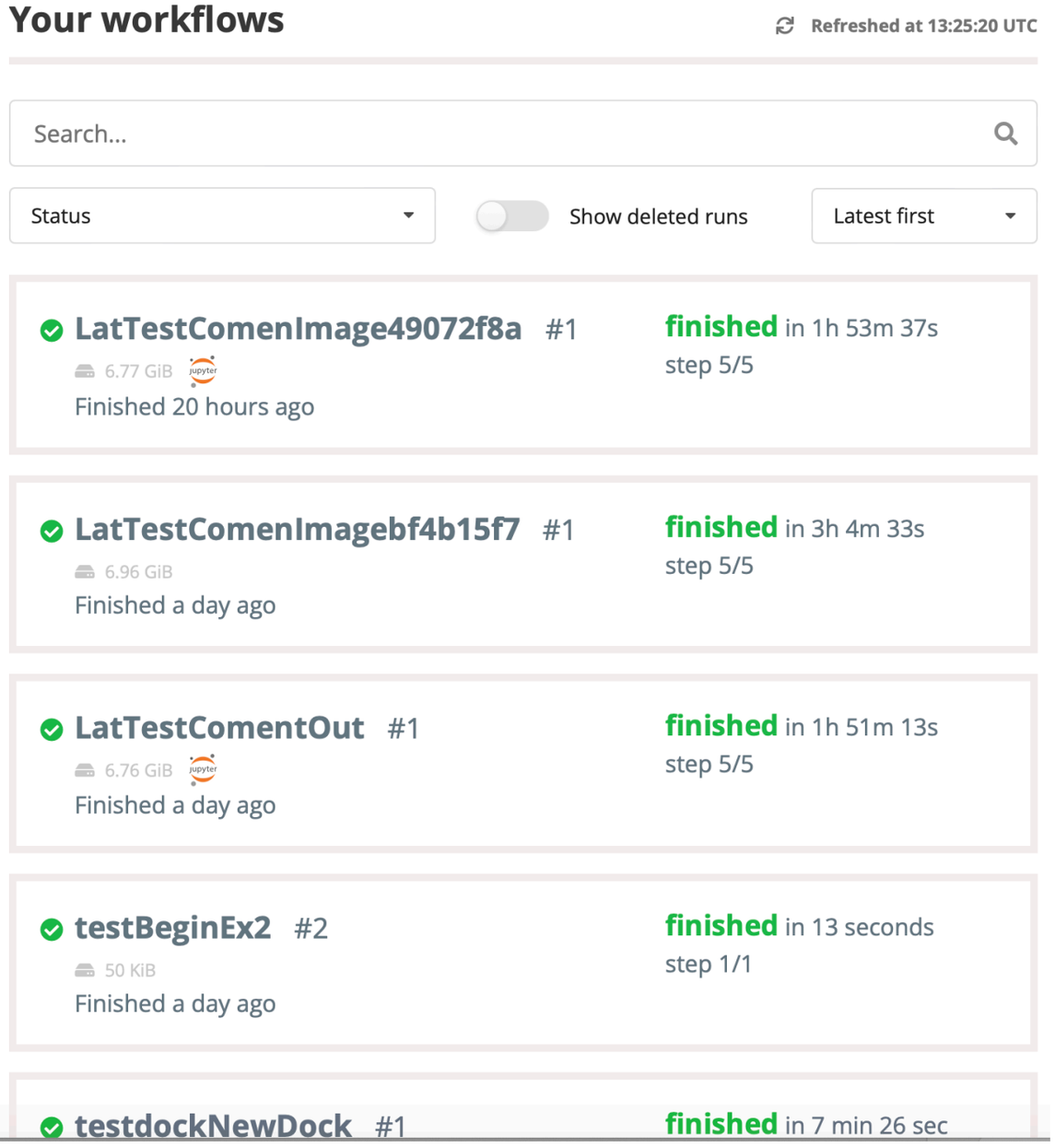
- Authors of the original workflow warned of this :

```
189
190   # 2.2 Spectral reconstruction [OPTIONAL, this takes a lot of computing time, so the output files are already included]
191   ./correlators_flow/spf_reconstruction/model_fitting/example_usage/spf_reconstruct.sh $BASEPATH_WORK_DATA NO $NPROC
```

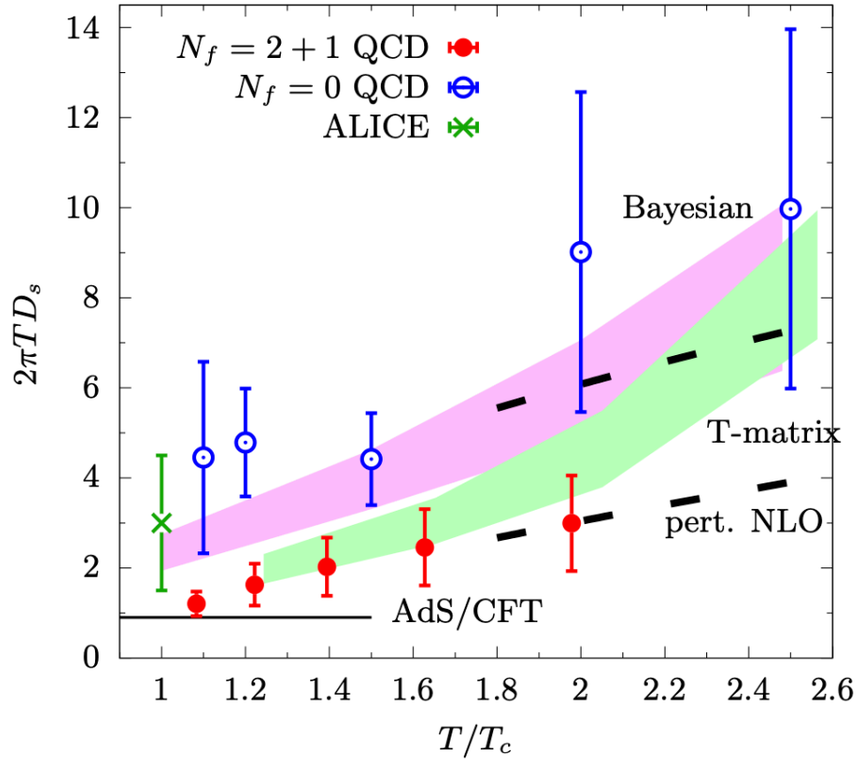
- Hence this part will be commented out for the next runs

Status on running the workflow on REANA : Results of workflow

- Results from some successful runs - different versions correspond to DockerImages
- Final plot of the paper successfully generated - however caveats remain



Some successfully generated figure from paper



Status on running the workflow on REANA : Testing the Jupyter Interface

jupyter

Files | Running | Clusters

Select items to perform actions on them.

0 /

- AnalysisToolbox
- correlators_flow
- figure_3_2piTD
- figures
- figureshisq_ms5_zeuthenFlow
- input
- output
- output_data
- download_data.sh
- HQDworkflow.sh
- reana.yaml

jupyter bootstr.py 20/05/2023 Logout

File Edit View Language Python

```python
#
# bootstr.py
#
# H. Sandmeyer, H.-T. Shu
#
# A parallelized bootstrap routine that can handle arbitrary return values of functions.
#
import numpy as np
from numpy.random import randint, normal, multivariate_normal
from latqcdtools.statistics.statistics import std_mean, std_dev, std_median, dev_by_dist
import concurrent.futures


def recurs_append(data, sample_data, axis, conf_axis, sample_size, same_rand_for_obs, i, my_seed):
    """ Recursive function to fill the sample. """

    np.random.seed(my_seed+i)
    if axis + 1 == conf_axis:
        numb_observe = len(data)
        if sample_size == 0:
            sample_sizes = [ len(j) for j in data ]
        else:
            sample_sizes = [sample_size]*len(data)

        if not same_rand_for_obs:
            randints = [randint(0, len(data[x]), size=sample_sizes[x]) for x in range(numb_observe)]
        else:
            tmp_rand = randint(0, len(data[0]), size=sample_sizes[0])
            randints = [tmp_rand]*numb_observe
        for x in range(numb_observe):
            sample_data.append(np.array(data[x])[randints[x]])
        return

    else:
        for j in range(len(data)):
            sample_data.append([])
            recurs_append(data[j], sample_data[j], axis + 1, conf_axis, sample_size, same_rand_for_obs, i, my_seed)

```

jupyter

Files | Running | Clusters

Select items to perform actions on them.

0 / **output_data** / **hisq_ms5_zeuthenFlow** / **EE** / **s064t24_b0777700**

- ..
- EE_imag_s064t24_b0777700_merged.npy
- EE_real_s064t24_b0777700_merged.npy
- flowtimes_s064t24_b0777700.dat
- n_datafiles_s064t24_b0777700.dat
- polyakov_imag_s064t24_b0777700_merged.npy
- polyakov_real_s064t24_b0777700_merged.npy

Some successfully generated output

jupyter flowtimes_s064t24_b0777700.dat Yesterday at 12:52

File Edit View Language

```
1  # flow times \tau_F for hisq_ms5_zeuthenFlow s064t24_b0777700
2  0.0000000000000000000e+00
3  1.7422199999999999999e-04
4  6.9688900000000000154e-04
5  1.5784709999999999911e-03
6  2.8296670000000000120e-03
7  4.4614219999999999610e-03
8  6.4849469999999999587e-03
9  8.9116979999999999148e-03
10 1.1753408999999999938e-02
```

Status on running the workflow on REANA : Conclusion and Outlook

- Open Issue 1 : How to successfully run the spectral function reconstruction scripts - right now this is commented out which leads to incomplete analysis

- Open Issue 2 : Too many files generated during the workflow (order ~5000) - not desirable for the cluster where workflow runs

  ➡ One solution is to install and import the (recently available) AnalysisTool box in docker instead of downloading the tar file.

AnalysisToolbox/latqcdtools/__init__.py	2023-05-20T10:07:38	669 Bytes
AnalysisToolbox/latqcdtools/math/math.py	2023-05-20T10:07:38	1.68 KiB
AnalysisToolbox/latqcdtools/math/optimize.py	2023-05-20T10:07:39	3.5 KiB
AnalysisToolbox/latqcdtools/math/spline.py	2023-05-20T10:07:39	5.21 KiB
AnalysisToolbox/latqcdtools/math/polynomials.py	2023-05-20T10:07:39	1.88 KiB
AnalysisToolbox/latqcdtools/math/num_int.py	2023-05-20T10:07:39	2.9 KiB
AnalysisToolbox/latqcdtools/math/SU3.py	2023-05-20T10:07:38	4.93 KiB

« ‹ 1 ••• 5358 **5359** 5360 ••• 5363 › »

- Open Issue 3 : Currently the intermediate figures get overwritten or deleted by the workflow - and only the final one gets saved.

Thank you for your attention !