Discussion about future directions for analysis

Jan Eysermans (Massachusetts Institute of Technology)

FCC Workshop, Annecy

January 30, 2024

FCC Analyses Survey

Purpose of the survey:

- General understanding of usage of the FCCAnalyses framework
- Feedback on what we can do better

19 did respond; honestly more than I was hoping for!

Thanks everyone that filled the questionnaire

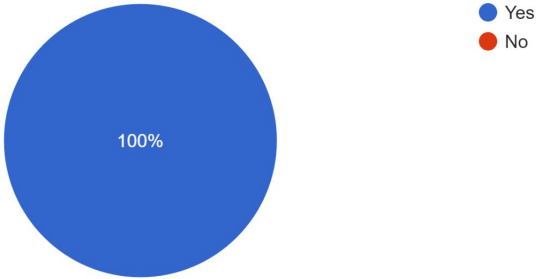
Let's go over the answers and guide the discussion

At the end go into more details, best practices, topics to discuss

Q1

Do you use the FCCAnalysis framework (https://github.com/HEP-FCC/FCCAnalyses) to perform your analysis for FCC-related studies?

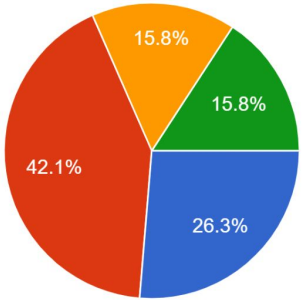
19 responses



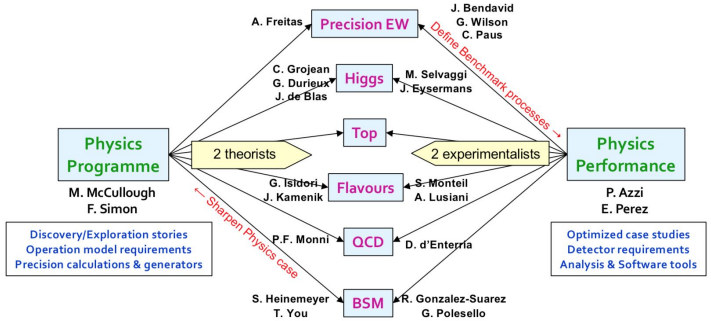
Q2

Do you run the analysis at LXPLUS at CERN?

19 responses



- Yes, and happy about it
- Yes, but I want more resources at CERN (e.g. too slow, need more space, ...)
- Yes, but I want to run on private/ institutional (e.g. own computer cluster/ HPC)
- No, I run locally/institutional resources



Seems almost everyone runs at CERN

More resources desired:

- Space resources: we have few TBs available for physics analyses, contact your physics performance conveners to ask for more space!
- Condor resources: only the CERN resources are available
- Any other resources you need?

Running outside CERN: currently not that trivial – see later for the prospects

Q: is there anyone running on a laptop/local machine, without cvmfs?

Q3

Do you use the Condor/batch pool (at CERN) to submit jobs, or do you use the analysis interactively?

19 responses

Quite heterogeneous usage of interactive/condor

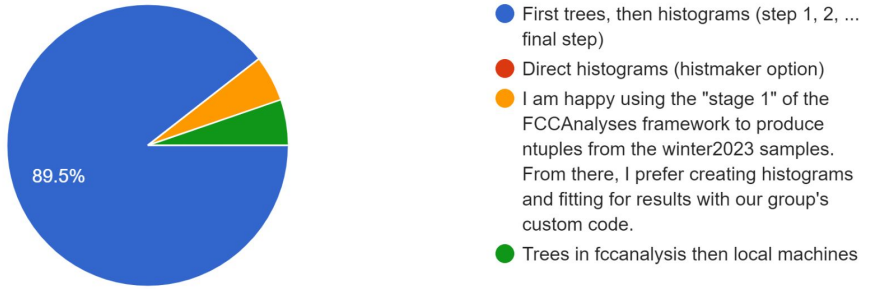
Q: What is the main reason of not using Condor?

- Interactively is good/fast enough
- Don’t want to use Condor
- Don’t know how to use it

Q4

When doing your analysis, do you first generate trees and then the histograms, or do you directly produce histograms in one step?

19 responses



The traditional analysis with intermediate Tree generation seems the most used

Q: are people aware of the histmaker option?

Q5

Would it be beneficial for you if the samples were remotely accessible (i.e. outside CERN), or that you can copy the necessary samples to your own storage?

18 responses

Clearly need to have samples remotely accessible

- We will provide instructions to copy desired samples + dicts to local space, and how to configure FCCAnalyses to change the paths of the samples

On long term will be solved by ILCDirac (where everything will be remotely accessible)

- But still need to have a way to just copy the RAW ROOT files

Q6

What is your experience with the MC samples and the EDM4HEP format?

19 responses

Often get the question: what is *FCCAnalyses::ReconstructedParticle* and what functions are behind? Are there more functions? Where can I find it? Info? Docs?

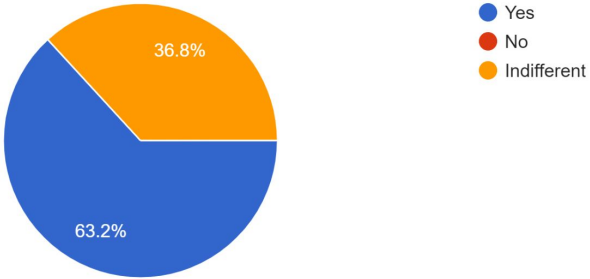
More documentation on the edm4hep structure is needed, in particular:

- How the collections are stored (#Muons, #Electrons, #Photons)
- Gen particles
- Relationships

Q7

Related to the question before, would it be useful to have flat ntuples in place for all the MC samples, that has a more understandable Tree structure (like the Nano format in CMS)?

19 responses



Related to previous question:

- Seems there is a desire for more flat ntuples where the physics objects are stored more easily
- More physics oriented, e.g. Px/Py/Pz → P/Theta/Phi
- Default jets/tagger collection?
- Can add pre-calculated physics variables (e.g. missing/visible mass, …)

We are still evaluating whether it is feasible/useful to have such flat ntuples

Q8

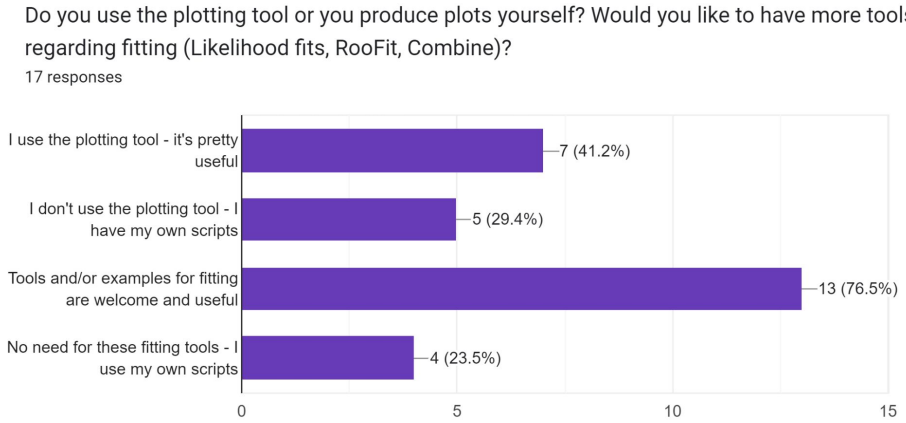
Do you find it useful to have the flavour tagging, vertexing and jet clustering in FCCAnalyses?

19 responses

Tools are appreciated

Likely more documentation is needed on what tools we have and how to use them

Q9

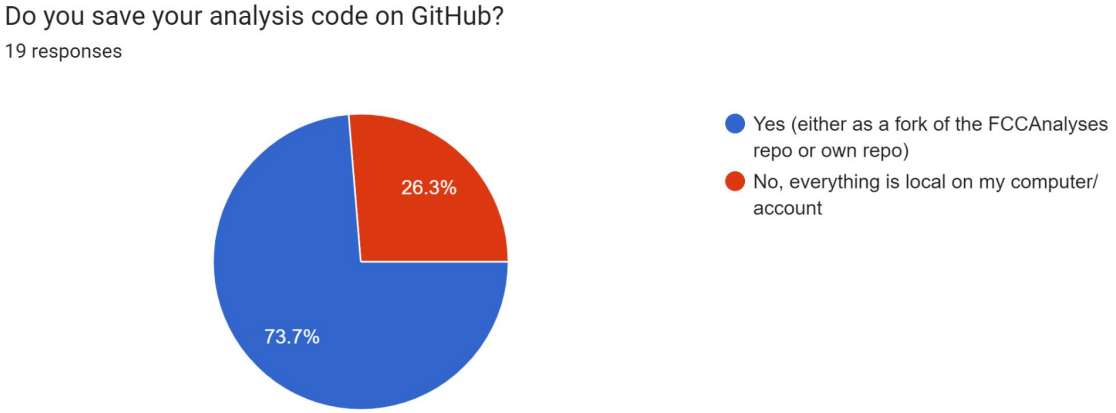


Plotting tool seems to be used

Fitting tools desired

See later for more info

Q10



Good practice!
But we need to have a good repo to store all the analysis files

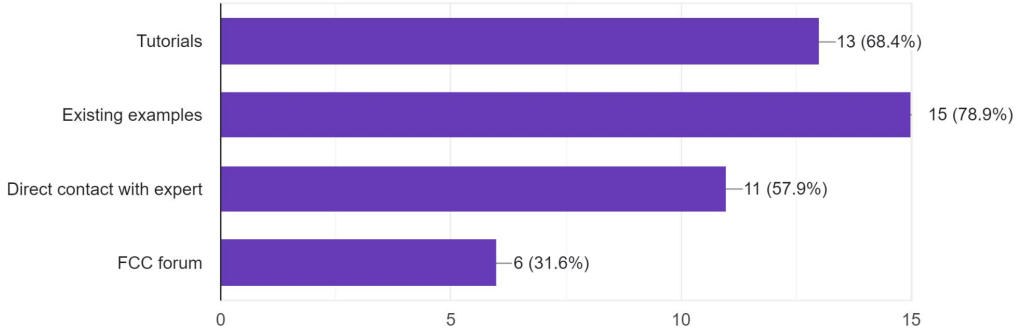
- Will work on an organized repository
- Example analyses will be stored accordingly

https://github.com/HEP-FCC/FCCeePhysicsPerformance

Q11

Where do you get the most information from to build your analysis?

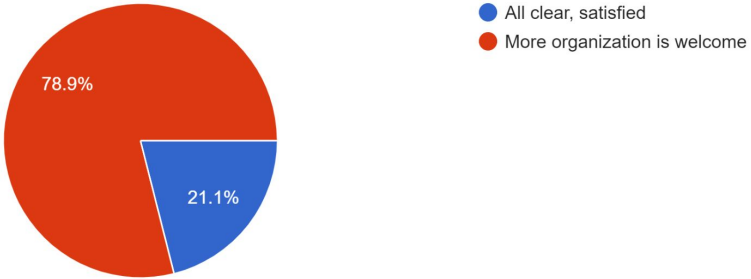
19 responses



Q12

Are you happy with the documentation and examples that we provide, or is it all a bit complicated/difficult to find?

19 responses



Work on our side!

Documentation

We are aware of the poor documentation :)

IDEA: we will prepare a new web page that contains all documentation centrally

- Provide “user setup”: source key4hep, create your Python file and ready to go
- Provide “advanced setup”: compile FCCAnalyses locally
- List of implemented functions: e.g. special computations of missing energy/mass, jet/hadron matching seems to be a popular request: maybe need to come up with a standard recipe
- List of tools that are implemented:
  - Jet clustering + explanations of what they do
  - Flavour tagging: how to use it
  - Tool for degrading detector parameters
  - ML support: working on support for TMVA/XGBoost/ONNX inference readers (see later)
- Examples and tutorials:
  - We have many examples and will polish them
  - Refurbish tutorials with more explanations
  - The idea is that newcomers pick an existing analysis close to theirs and start from this implementation

Code Versioning

Often (including myself) after some time of inactivity, I source my code and find that the setup is broken

- Need to recompile or just start from scratch

This naturally calls for code versioning, or tagging, to have stable releases

- We are still investigating the best way of providing stable releases
- Not so easy as the code is also rapidly developing with new features

The idea at least is that all provided examples will be added to the CI, and tested when changes are made to the FCCAnalyses framework

- In this way we try to catch as much as possible bugs

Machine Learning Inference

High demand for ML interfaces

- We have support for ONNX, but need to wrap it nicely in the framework
- Add support for basic BDT/MVA training using ROOT/XGBoost

We will provide basic examples to create to train, test and apply using XGBoost

Any other features needed?

Running Analysis outside CERN

Seems most of the users run at CERN, but there is desire to run outside CERN

- This will increase with more engaging person power
- Especially people outside Europe

Example at MIT

- Can’t read EOS nor via XrootD
- We copied the entire Winter2023 production to local storage (took some time
- Run everything on our cluster

Todo list:

- We will provide asap documentation on how to copy (part of) the samples to local storage
- FCCAnalyses need to be flexible to run on different Condor clusters (or Slurm)

Samples and Normalization

Desire to run also on locally produced samples

- Need to add flexibility to add local processes with specific path and cross-section

The final product of each analysis are histograms, normalized to

- Luminosity → input user. Where? Plotting? Stages? Histos?
- Cross-section / k-factor → fixed in db, should be able to change by user if desired
- Number of events / sum of weights → should be calculated on the fly

Sum of weights currently is not calculated, but taken from the prodDict (fixed number)

- Not flexible: suppose I want to run on a subset of MC files
- Will change that ASAP: calculate sumw on the fly
- Need to keep track of this number(s) throughout the different stages (stepX, final)

Plotting and Fitting Tools

Plotting tool seems to be used

- The code behind is outdated and we will rewrite it from scratch
- Functionality will be similar, maybe the interface will be slightly different
- Any suggestions/functionalities needed?

Fitting tools

- Promotion of Combine (CMS tool), based on RooFit
- https://cms-analysis.github.io/HiggsAnalysis-CombinedLimit/
- It can handle both binned and unbinned/parametric fits
- We have plenty of existing examples using Combine already:
  - Higgs mass and cross-section: both parametric (mass) and binned (xsec)
  - Hadronic couplings/mumu/gaga: binned
- Will document these examples

Running on Condor

Need to provide guidelines

- When to use Condor
- How to use Condor, and how it is different than running interactively
- Splitting of jobs – practical guidelines, automatic splitting?

Examples/Best practices on how to run

Example 1: Higgs mass analysis (link to file)

- Samples all Higgs signals, WW/ZZ/Zg backgrounds → WW huge sample!
- Using the “histmaker” mode: directly produce the histograms (omit intermediate steps)
- Run fully multithreaded on a 256-thread machine
- Takes 1 hour to run the entire muon and electron analysis

Example 2: Higgs couplings

- Samples all Higgs signals, WW/ZZ/Zg backgrounds → WW huge sample!
- Need for jet clustering and flavour tagging: slow inference, want to do it once
- Therefore multi-stage approach better:
  - Step 1: create trees with loose selection + clustering/flavor inference – “static”
  - Final step: desired histograms – “variable”

Conclusions

- The way to run depends on the analysis and available resources
- Get in touch with us to discuss your best approach – we’re happy to help

Monte Carlo Samples

Didn’t touch it in the survey, but any feedback on Monte Carlo samples?

- Need to have a workflow to make private samples + analyze them using FCCAnalyses?
- Happy about the generators, number of events, ….?
- Anything else?