10.1088/1748-0221/13/12/P12018

A GPU-based framework for multi-variate analysis in particle physics

25 Jan 2019, 12:10 (CET) @503-1-001, 12'+3'

Xuefeng Ding[1,2]

1. Gran Sasso Science Institute, L'Aquila, Italy
2. INFN Sezione di Milano, Milan, Italy

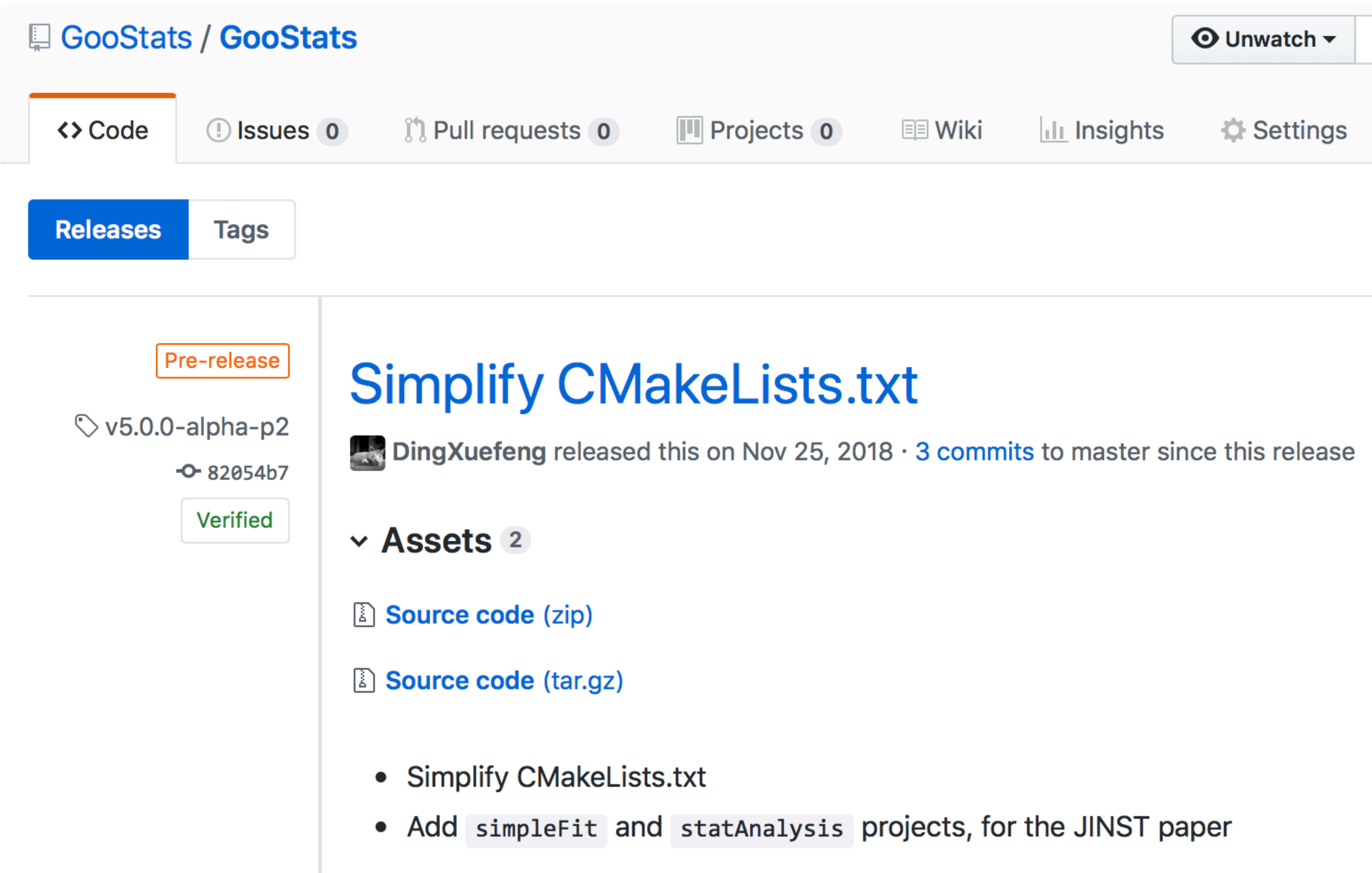
PHYSTAT-nu 2019
@ CERN, Geneva, Switzerland 25 January 2019

Outline

- The GooStats framework
- Motivation: Borexino spectrum fit
- Package design
- Systematic uncertainty estimation
- Conclusion

GooStats hosted on GitHub

https://github.com/GooStats/GooStats.git



What and Why?

- A convenient GPU multivariate analysis framework
- Easy spectrum fitting

```
./fit toyMC.cfg out exposure=500
```

```
FIT PARAMETERS
 gaus = 14.14 #pm 0.21 (day#timeskt)^{-1}
 fbkg = 0.588 #pm 0.016 (day#timeskt)^{-1}
 sdn = 0.991 #pm 0.014
 ^{1}v = 0  [Fixed]
 sigmaT = 0  [Fixed]
 gaus Epeak = 5.000 #pm 0.014  [p] 4.90 #pm 0.50

 chi^2                          = 112.7
 chi^2/N-DOF                    = 1.1739
 p-value                        = 0.117
 Minimized -2Ln(Likelihood)     = 714.53
 Likelihood p-value             = 0.540 ± 0.050
```

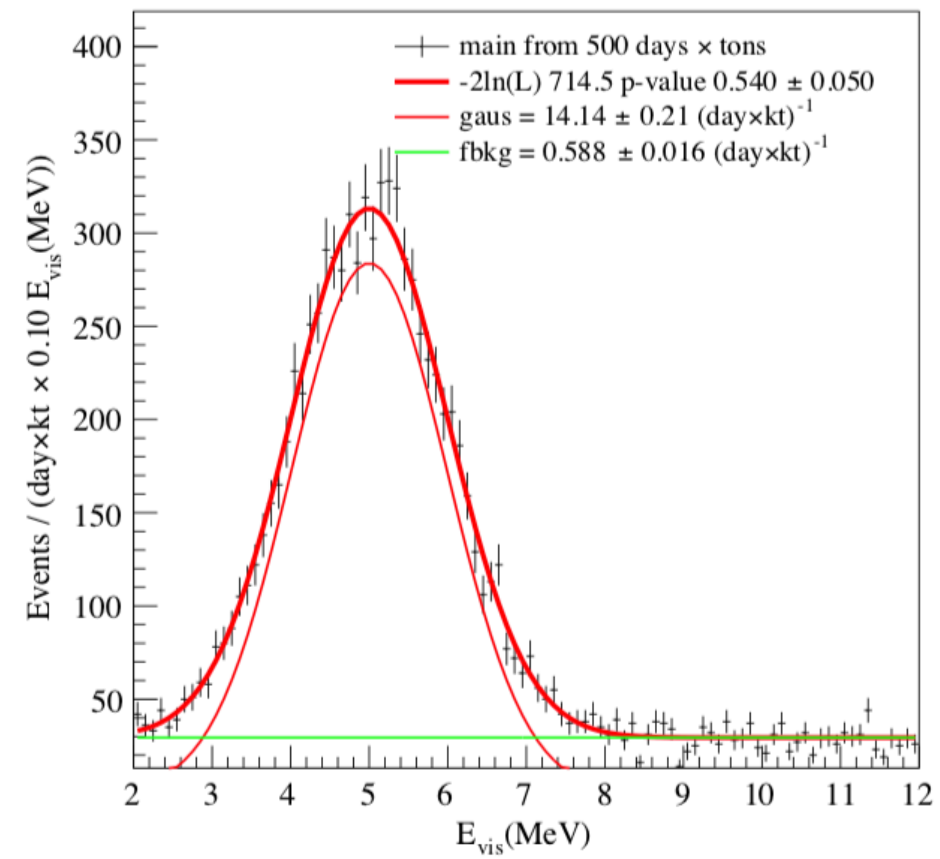


Figure 7. Left: screen shot of fit result summary. Right: produced figure in the file format of pdf.

GooStats: performance

- Speed up: around 200 times

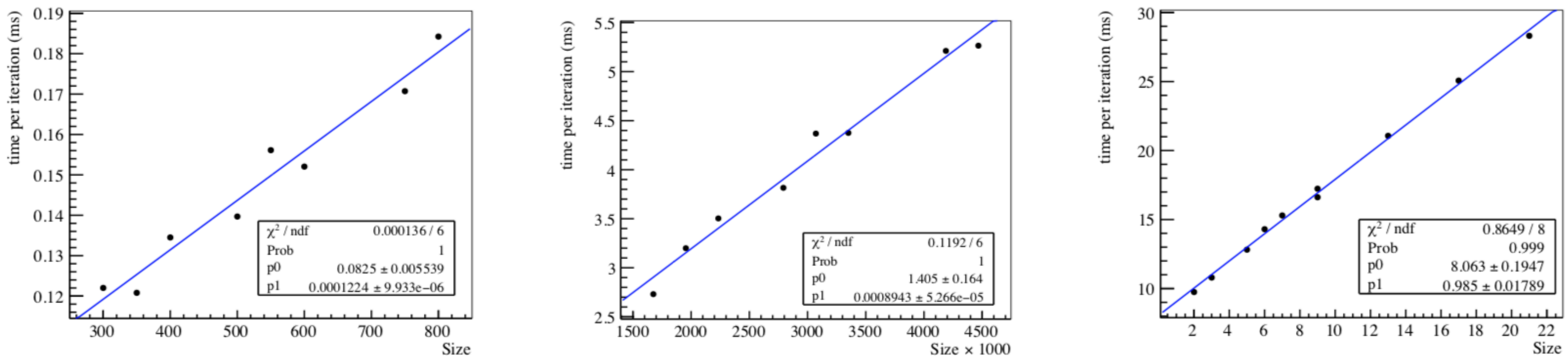
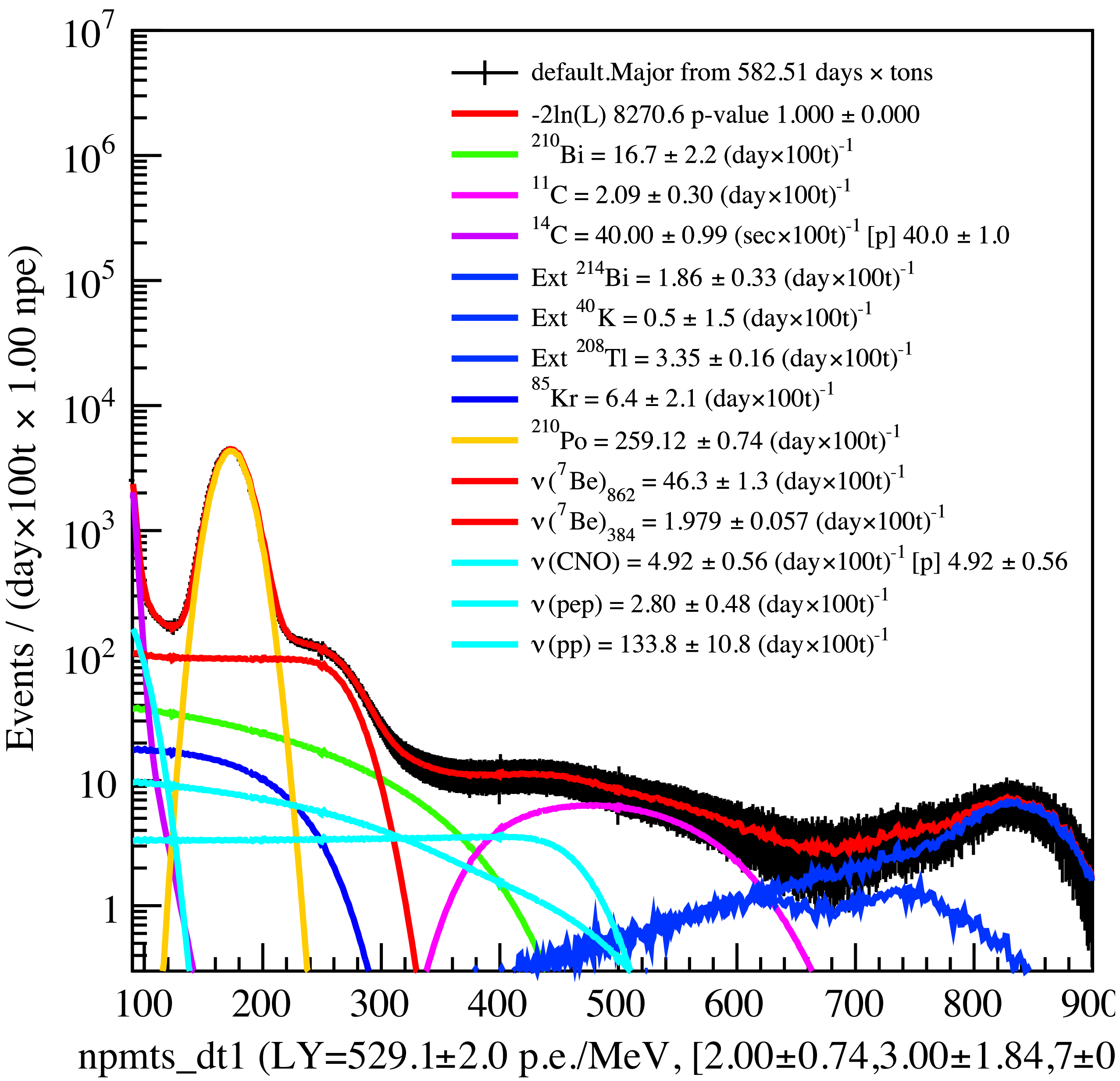


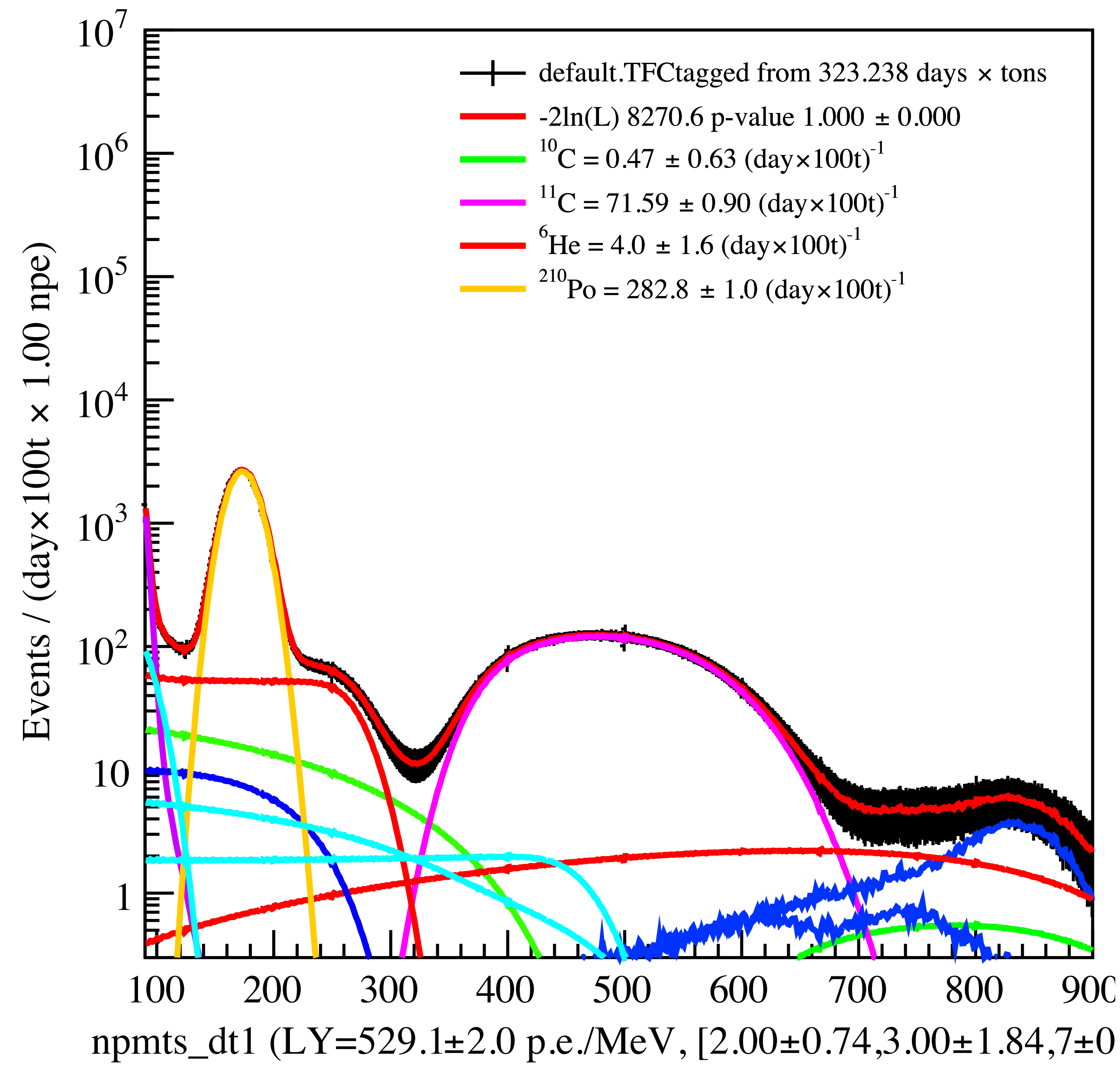
Table 1. Comparison of fitting time between `GooStats` and original software used by the Borexino collaboration. T_{tot}: total execution time. N: the number of iterations taken to converge in `MINUIT`. T_{it}: average execution time per iteration. Speed up: $T_{\mathrm{it}}(\mathrm{CPU})/T_{\mathrm{it}}(\mathrm{GPU})$.

Type	CPU AMD Opteron(TM) Processor 6238			GPU nVidia Tesla K20m			
Size	T_{tot} (s)	N	T_{it} (ms)	T_{tot} (s)	N	T_{it} (ms)	speed up
400	27.6	1128	24.4	0.181	1346	0.135	181
350	29.4	1331	22.1	0.156	1294	0.121	183
300	22.5	1239	18.2	0.243	1995	0.122	149

Motivation: Borexino spectrum fit

- Motivation: spectral-fitting tool for Borexino analysis



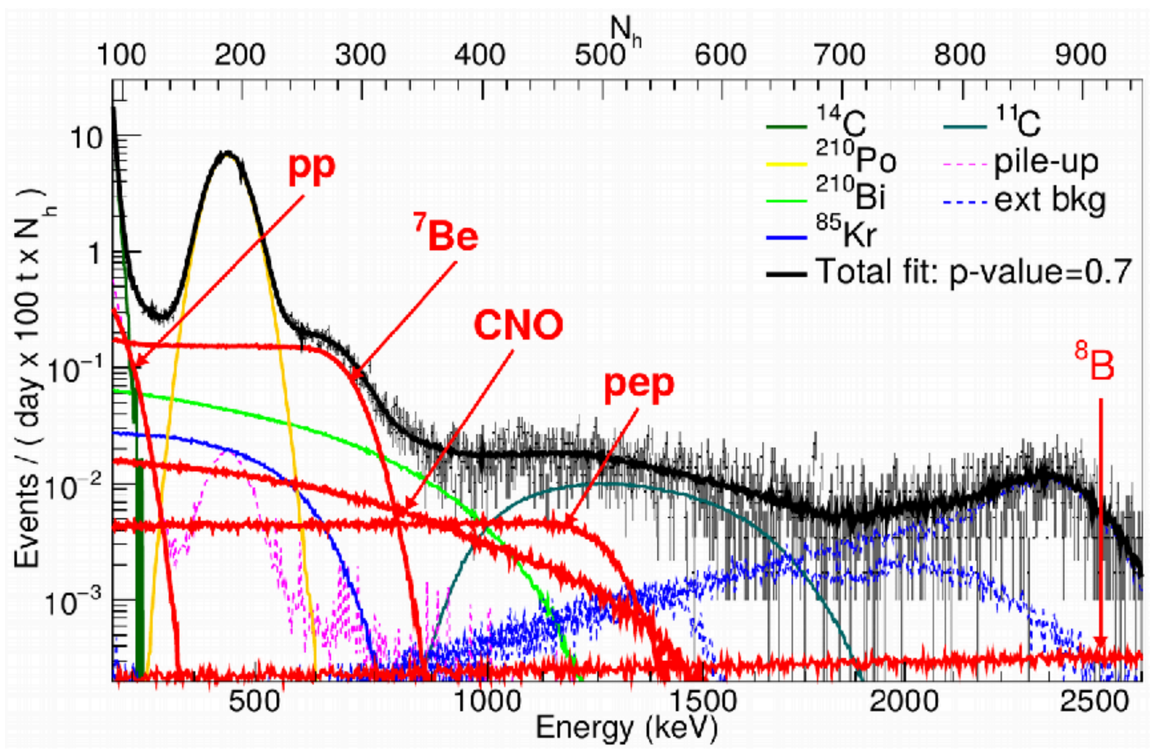


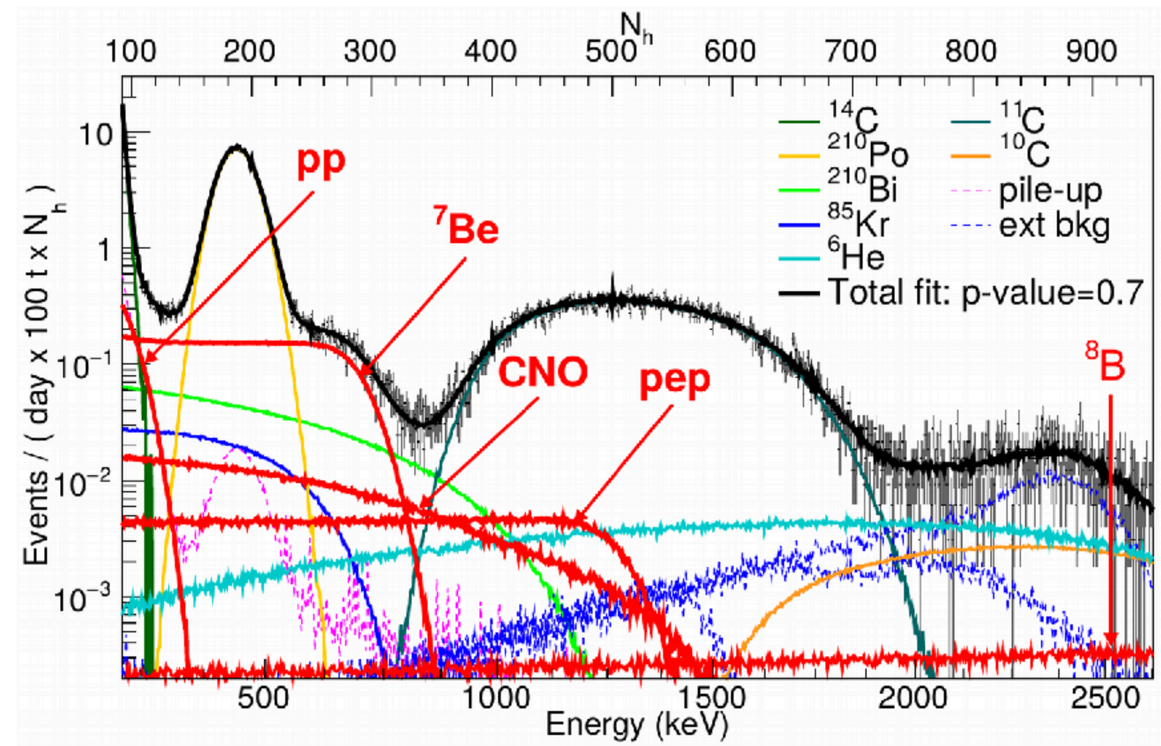
Analytical Response Function

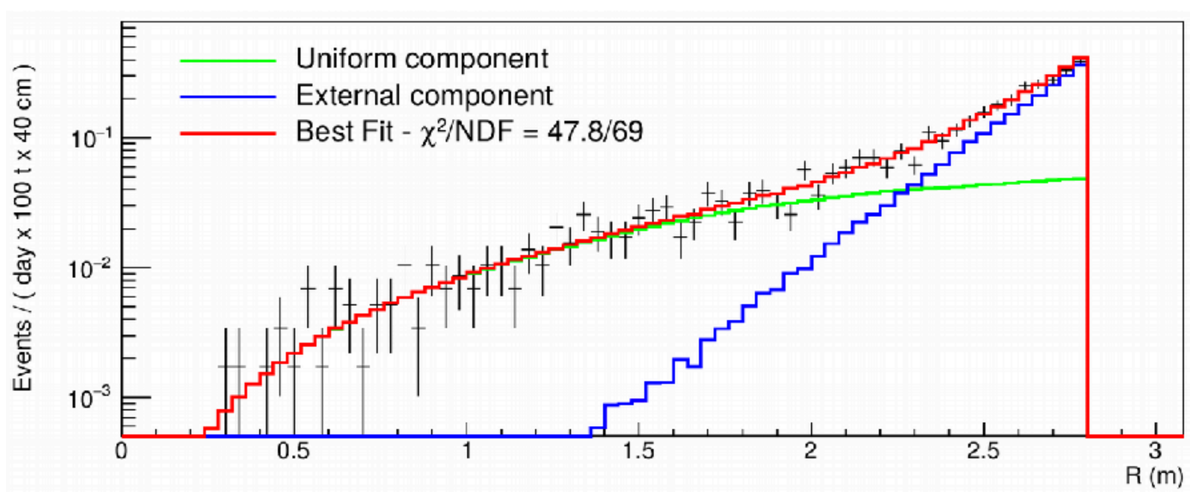
$$M : f(E) \mapsto g(\text{charge}) = \int_0^{E_{\text{end}}} \mathrm{d}E \cdot f(E) \cdot \mathrm{RPF}\left[\text{charge}; \mu(E), \text{var}(\mu)\right]$$

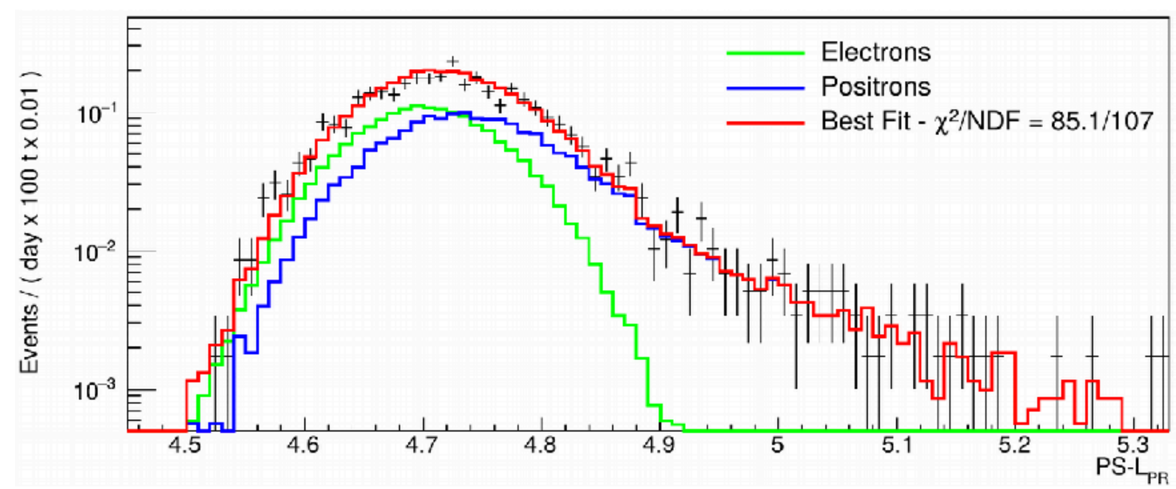
- Analytical shape of spectrum of mono-energetic events
  - **Momentum based approximation**
  - Match the average (energy scale + non-linearity model)
  - Match the variance (energy resolution model)
  - … (—> simplified)
  - More: “Mask”, “pile-up” etc…
- We can simplify because
  - Borexino response is simple: small FV in center, low energies => no irregular tail
  - We are not sensitive.. => small systematics
  - **Fit full MC to get the bias introduced in simplification**

Multivariate Likelihood (Davini method)







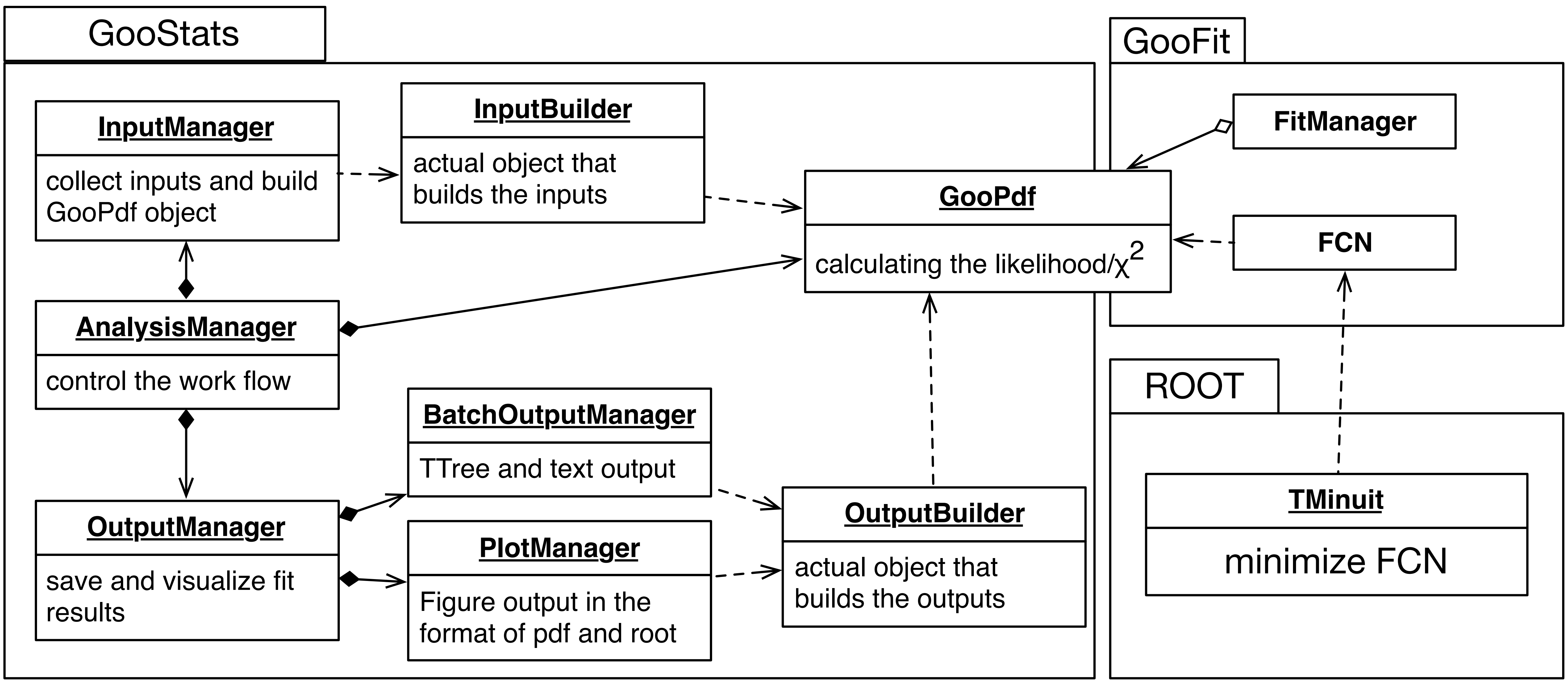


$$\mathcal{L}_{\mathrm{MV}}\left(\vec{\theta}\right)=\mathcal{L}_{\mathrm{TFC-sub}}\left(\vec{\theta}\right)\cdot\mathcal{L}_{\mathrm{TFC-tagged}}\left(\vec{\theta}\right)\cdot\mathcal{L}_{\mathrm{RD}}\left(\vec{\theta}\right)\cdot\mathcal{L}_{\mathrm{PS}}\left(\vec{\theta}\right)$$

- Scaling factor introduced to remove bias.

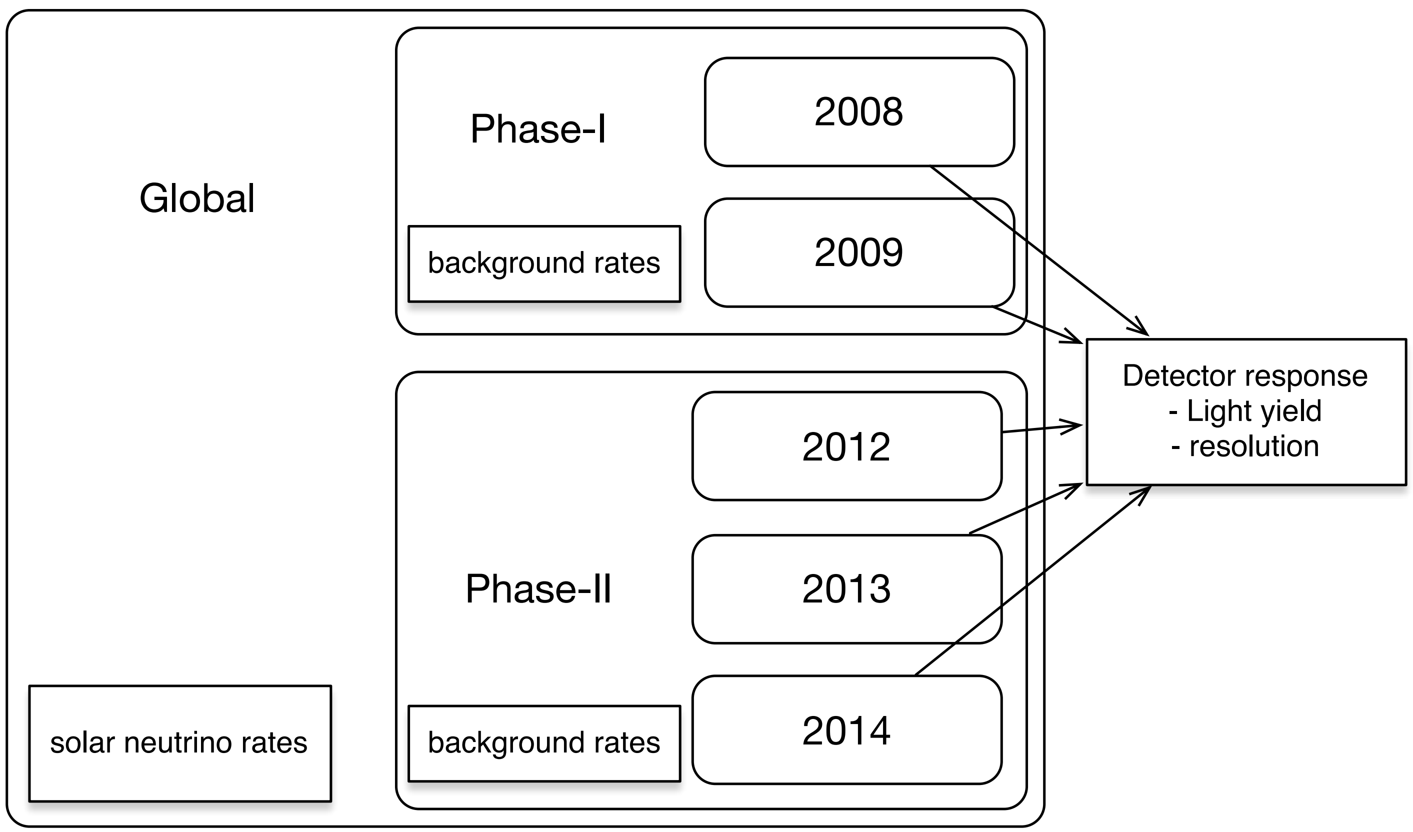
Package Design

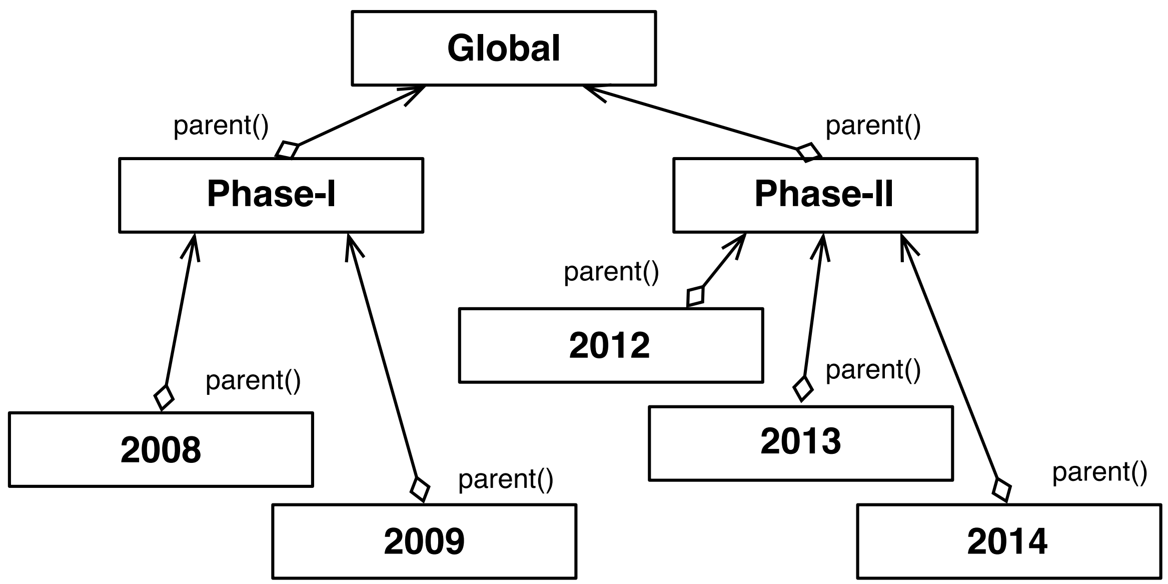
- **Middle Layer** between **GooFit** and **User module**



Parameter Synchronization

- Synchronize fit parameters in joint analysis
- Supported by tree type internal data structure





TTree Output

- All fit parameters saved in TTree output automatically
- More quantities can be added as lambda functions in the output builder

```
root [1] .ls
TFile**          test_tree.root
 TFile*          test_tree.root
  KEY: TTree     fit_result;1     Fit result of GooStats
root [2] fit_result->Show(0)
======> EVENT:0
 default.NReactor = 3.02987
 default.NReactor_err = 0.0159146
 default.Nbkg     = 1.02405
 default.Nbkg_err = 0.0162459
 default.U235     = 0.5
 default.U235_err = 0
 default.U238     = 0.2
 default.U238_err = 0
 default.Pu239    = 0.1
 default.Pu239_err = 0
 default.U241     = 0.2
 default.U241_err = 0
 default.LY       = 1300
 default.LY_err   = 0
 default.qc1      = 2.78788
 default.qc1_err = 0
 default.qc2      = -0.528003
 default.qc2_err = 0
 default.v1       = 0.3
 default.v1_err   = 0
 default.vT       = 5
 default.vT_err   = 0
 default.Reactor_dEvis = 1078.28
 default.Reactor_dEvis_err = 13.2291
 chi2             = 390.448
 NDF              = 397
 likelihood       = 1883.96
```

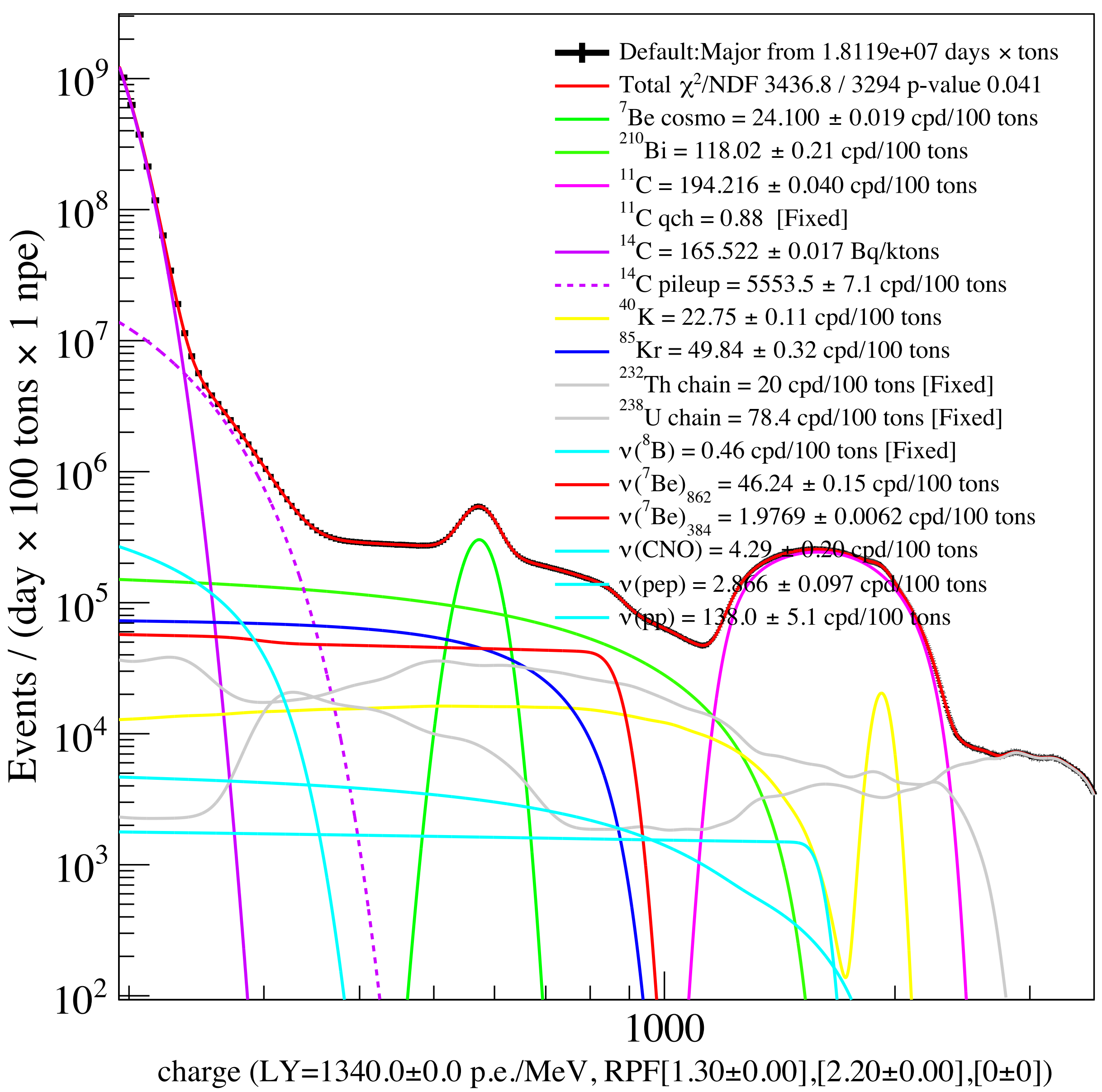
Statistical analysis modules

- Adding tasks of statistical analysis by registering new modules inheriting abstract classes
- A few pre-installedt modules: ScanPar, DiscoveryTest..

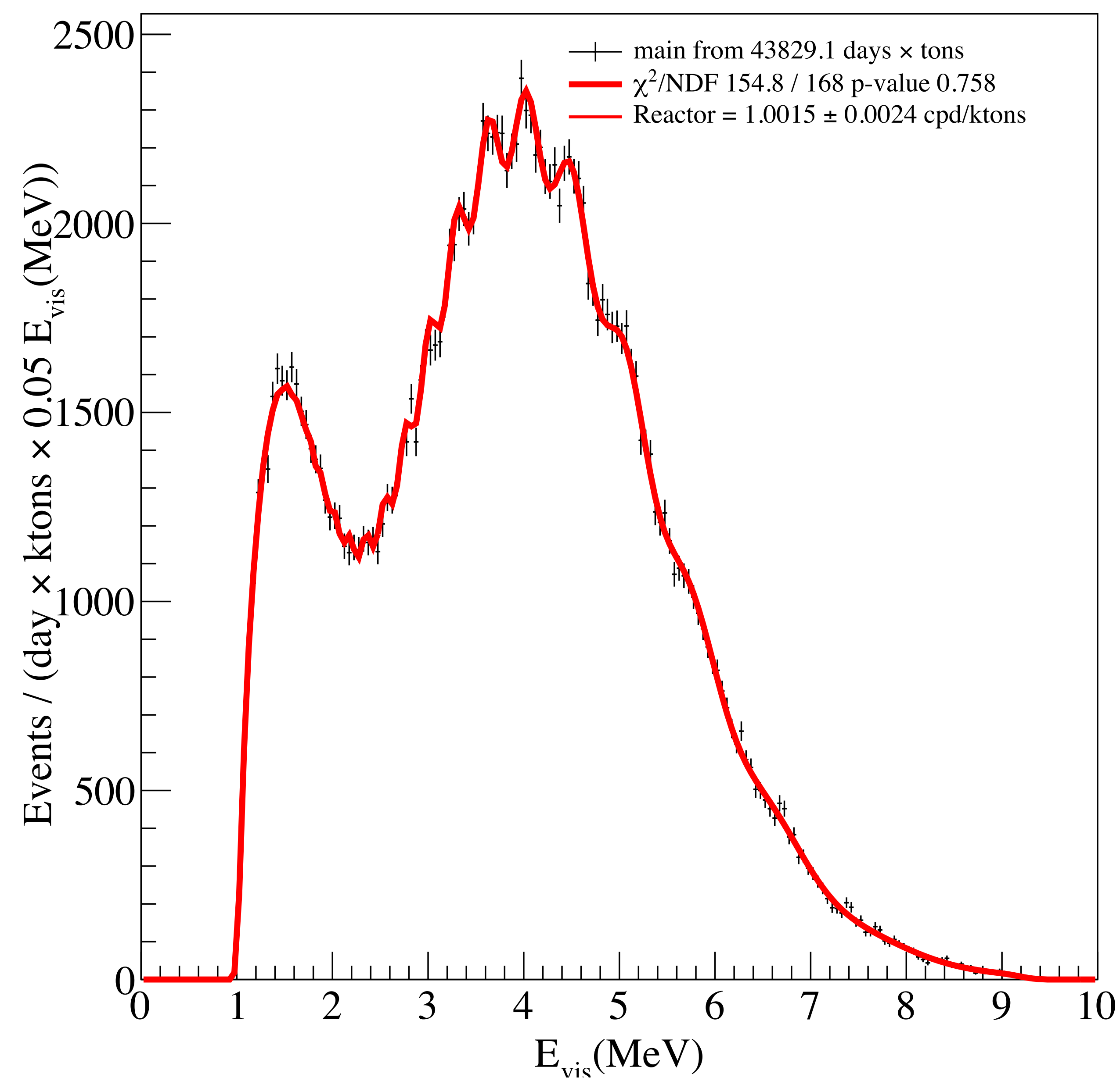
```
PrepareData *data = new PrepareData();
SimpleFit *fit = new SimpleFit();
DiscoveryTest *discovery = new DiscoveryTest();

ana->registerModule(inputManager);
ana->registerModule(data);
ana->registerModule(fit);
ana->registerModule(discovery);
ana->registerModule(outManager);
```

Physics modules

Solar Neutrinos



JUNO style



Two ways of estimating σ_{sys}

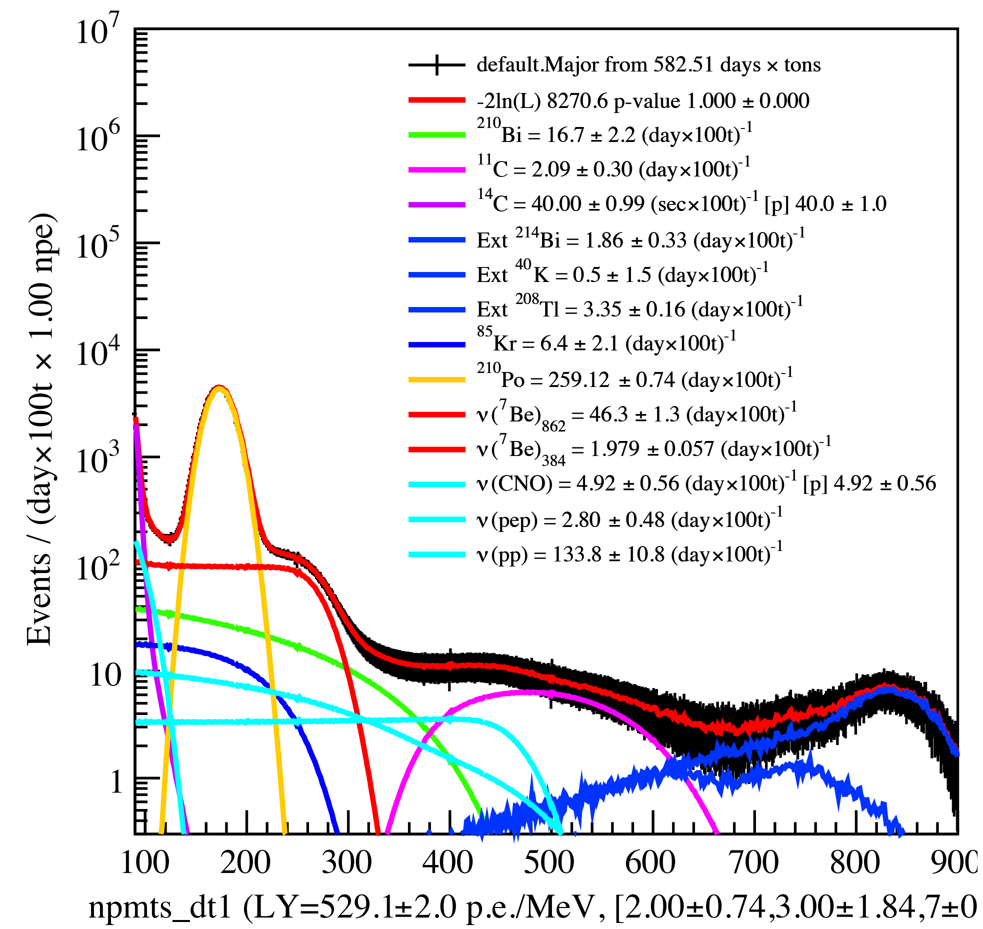
- create an ensemble of models according to experiment precision
  - Fit the data with varying models through their coordinates
  - Generate pseudo-experiments with varying models and take the width of distribution of best-fit

Method 1: fit with varying models

- **Full MC** during each iteration of the fit, vary **kb / absorption length spectrum etc.** **re-simulate and produce new pdf on the fly** —> when one day computer is fast enough
    - ~200, 000 CPU x years per fit
- **Semi-analytical:** analytical non-linearity model + response Matrix
    - ~30 minutes per fit
- **Full analytical**
    - ~2 hours per fit

Problem with scaling response matrix

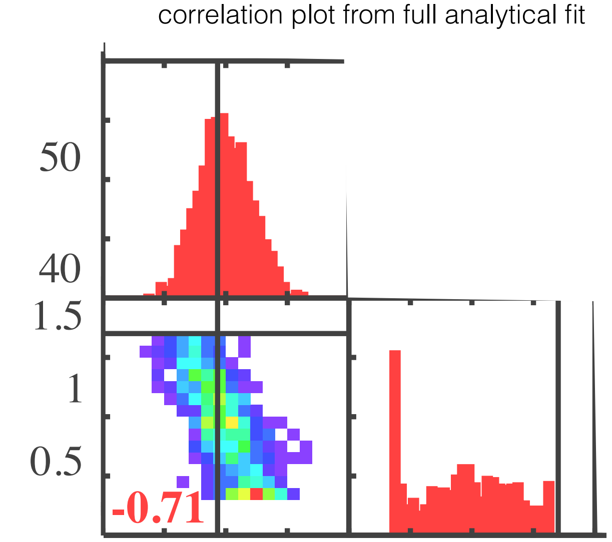
- Systematic uncertainties of LY can be included by scaling the response matrix
- This could be dangerous if the interested parameters is correlated with the resolution parameters
- When LY is scaled, the resolution should also be changed



$\nu(^7Be)$
48.5 (-2.1 +2.3)
inj 48.00 ± 0.07

β-RPF(0)
0.9 (-0.4 +0.3)
inj 1.49 ± 0.01

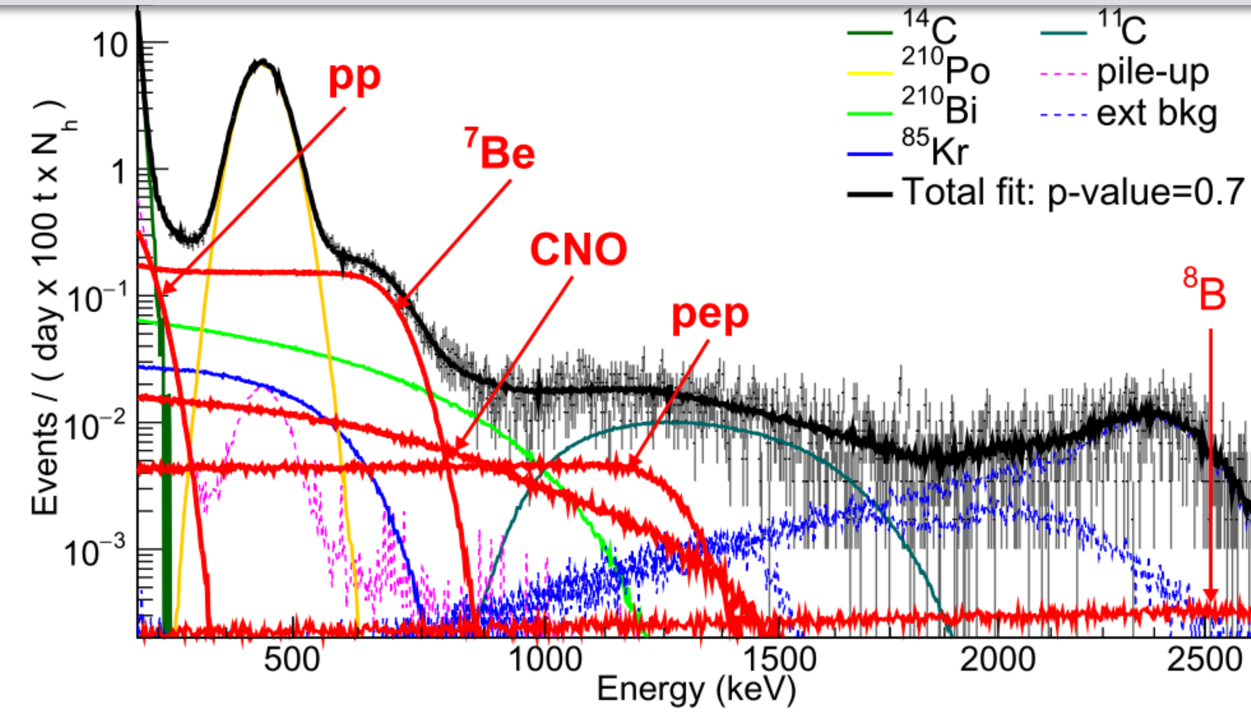
correlation plot from full analytical fit



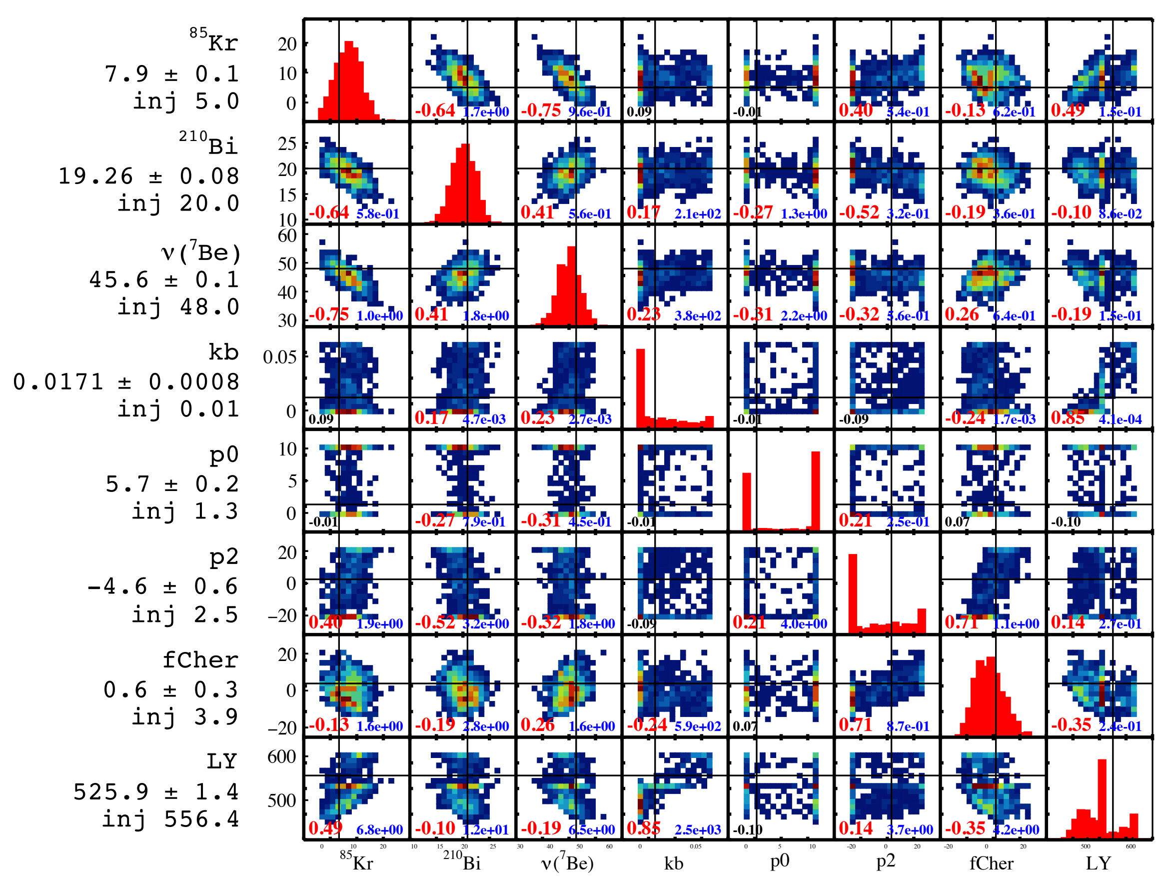
Correlation with NL/resolution

- **Using full analytical response function, we can see the correlation with detector responses**

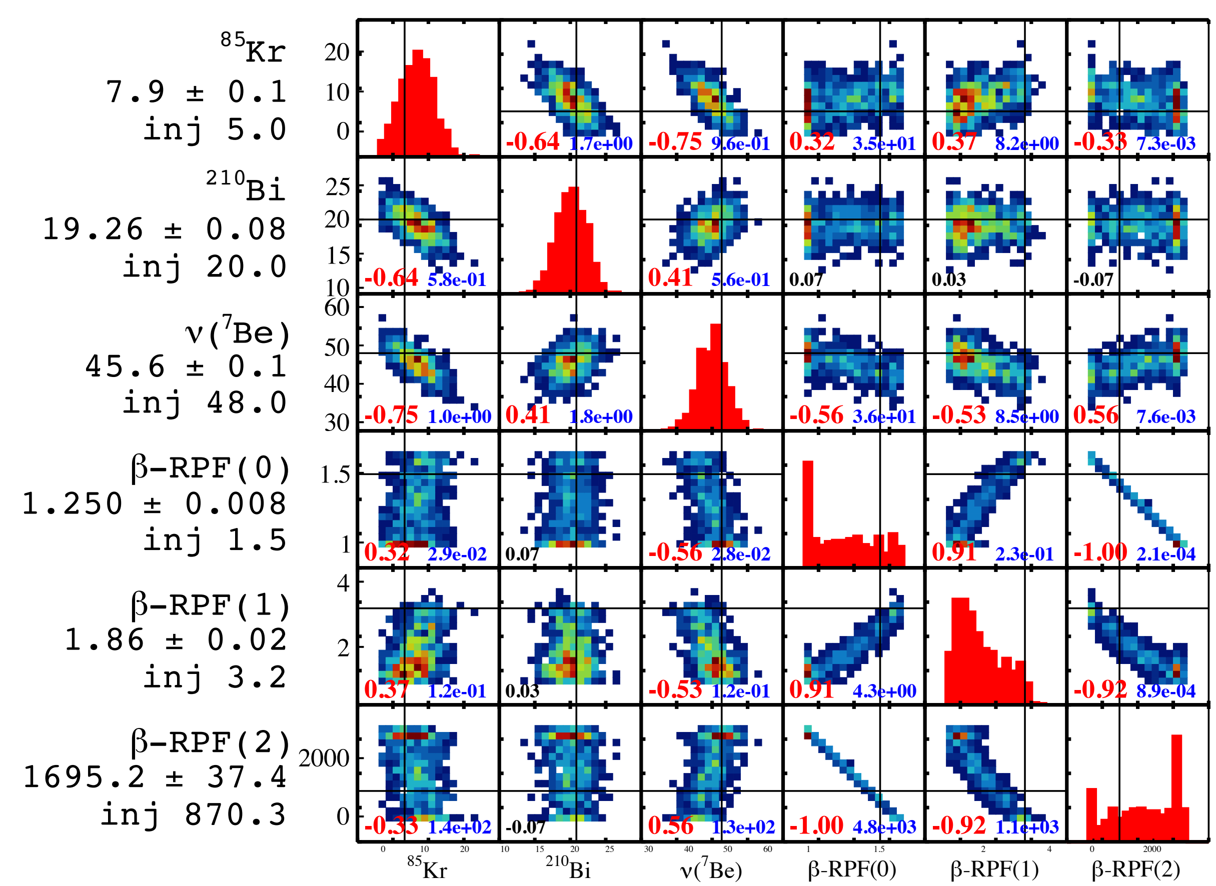
[1] M. Agostini *et al.*, “First Simultaneous Precision Spectroscopy of pp, ^{7}Be, and pep Solar Neutrinos with Borexino Phase-II,” pp. 1–8, Jul. 2017.



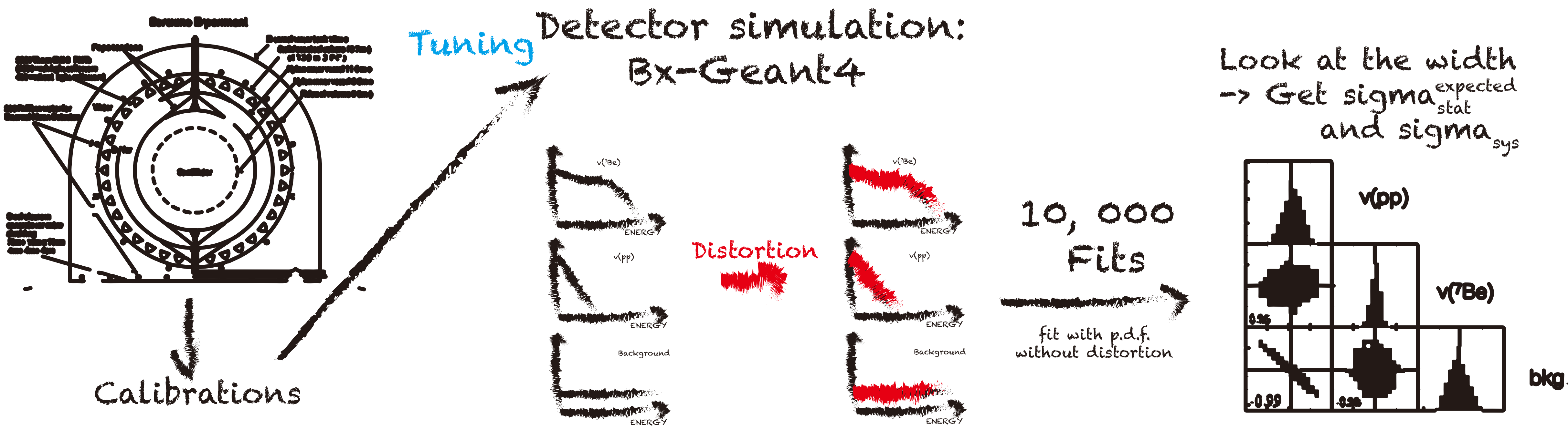
correlation with NL



correlation with res



Method 2: Monte Carlo method



- pseudo-experiment spectra without distortion —> **statistical sensitivity**
- pseudo-experiment spectra with distortion —> **statistical** **systematic uncertainty**

Conclusion

- GooStats, a GPU based multivariate analysis framework is introduced.
- Statistical analysis module is easy to be implemented as needed. TTree/figure output provided.
- Full analytical response function has advantage that it treated the NL and resolution in a coherent way when evaluating the systematic uncertainties
- The systematic uncertainties can also be evaluated using Monte Carlo method by looking at change of width of best fit with/without distortion