Template fits: TRExFitter et al.

Alexander Held[1,2,3]

[1] *University of British Columbia*
[2] *TRIUMF*
[3] *starting at New York University soon!*

June 19, 2019

Statistical analysis frameworks

- **Template fit user story:** https://github.com/iris-hep/as-user-facing/blob/master/template-fit.md

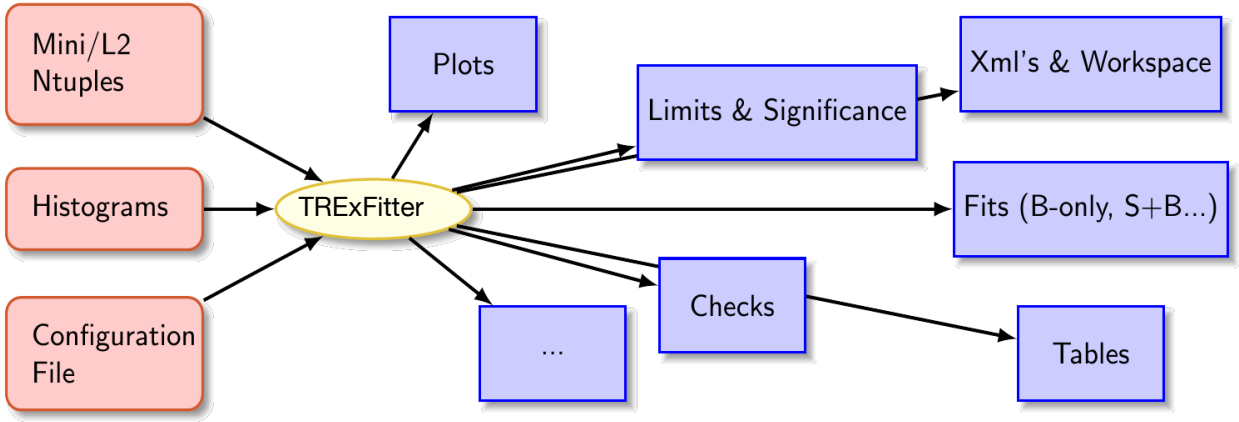
- **TRExFitter** developed from scripts during Run-1 of the LHC and developed into a **profile likelihood framework**
  - ATLAS-internal framework, https://gitlab.cern.ch/TRExStats/TRExFitter

- **Multiple additional frameworks** built with **RooStats/RooFit/HistFactory** are in use, for example
  - **HistFitter** in ATLAS (Eur.Phys.J. C75 (2015) 153)
  - **CMS Combine** https://github.com/cms-analysis/HiggsAnalysis-CombinedLimit
  - **Different frameworks** generally have **similar approaches,** shaped by conventions in the groups using them

- **General workflow:**
  - **Declare a fit model**, and **provide input** ntuples or histograms
  - Framework **builds workspace** for statistical analysis (this needs to be parallelized and fast!), and **performs fits**
  - Framework provides **diagnostics tools** and allows to **easily adjust the fit model** to study the fit
    - This is the heart of the framework
  - Also provides **figures, tables, etc**. for publication

The TRExFitter project

- **TRExFitter:** framework used by many ATLAS analyses for **statistical inference via profile likelihood fits**
  - Users provide inputs in various formats and a configuration file steering the framework
  - **TRExFitter** does the rest:
    - Produces **RooFit** workspaces, performs fits on them, interfaces with widely used **RooStats** macros (e.g. for limits)
    - Generates plots, tables, … to document the fit and help analysers study it in detail
    - **Crucial to success:** simplify validation studies, and provide as many diagnostic tools as possible



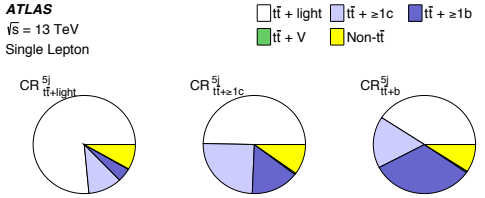
The configuration file

- **The configuration file steers the framework:**
  - **Plain text file** with simple structure
  - Divided into ***blocks***, separated by blank lines
  - The **Job** block defines general options
  - The **Fit** block specifies details of the fit model
  - **Region** blocks define the distributions considered in the fit
    - Called channels in HistFactory
    - Specify variables and cuts
  - **Sample** blocks define the samples considered
  - **Systematic** blocks specify systematic uncertainties
- **TRExFitter** builds histograms and a **HistFactory workspace** using this specification

```
Job: "ttH2015_ljets"                                   <- comments after %
  POI: "SigXsecOverSM" % ttH signal strength
  ReadFrom: NTUP
  NtuplePaths: "/afs/cern.ch/work/p/pinamont/public/flatNtup_TTHBB_test/"
  MCweight: "FinalWeight*weight_leptonSF*weight_bTagSF"
  ...

Fit: "fit"
  FitType: SPLUSB
  FitRegion: CRSR
  UseMinos: SigXsecOverSM
  ...

Region: "ljets_HThad_4j2b"                             <- exemplary Region block
  Type: CONTROL
  Variable: "HhadT_jets/1e3",50,100,1100
  VariableTitle: "H_{T}^{had} [GeV]"
  Label: "4 j, 2 b"
  ...

Sample: "Data"
  Title: "Data 2015"
  Type: DATA
  NtupleFile: "data"
  ...

Systematic: "tt_Shower"
  Title: "ttbar PS and hadronization"
  Type: HISTO
  Samples: ttbar
  NtupleFileUp: "ttbar_pyt8"
  ...
```

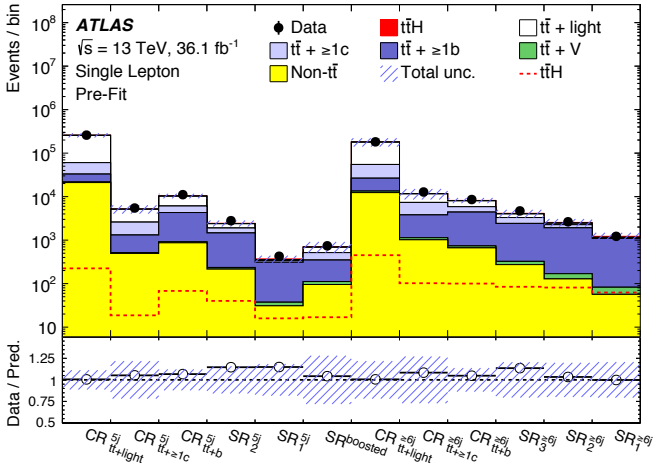
Input histogram creation and first plots

- **Workspaces** are **built from histograms**, and thousands of these are needed
  - One histogram per region (channel), per sample, per systematic variation
  - Framework takes other histograms or ntuples as inputs
  - This step needs to be parallelizable, or can take hours to days



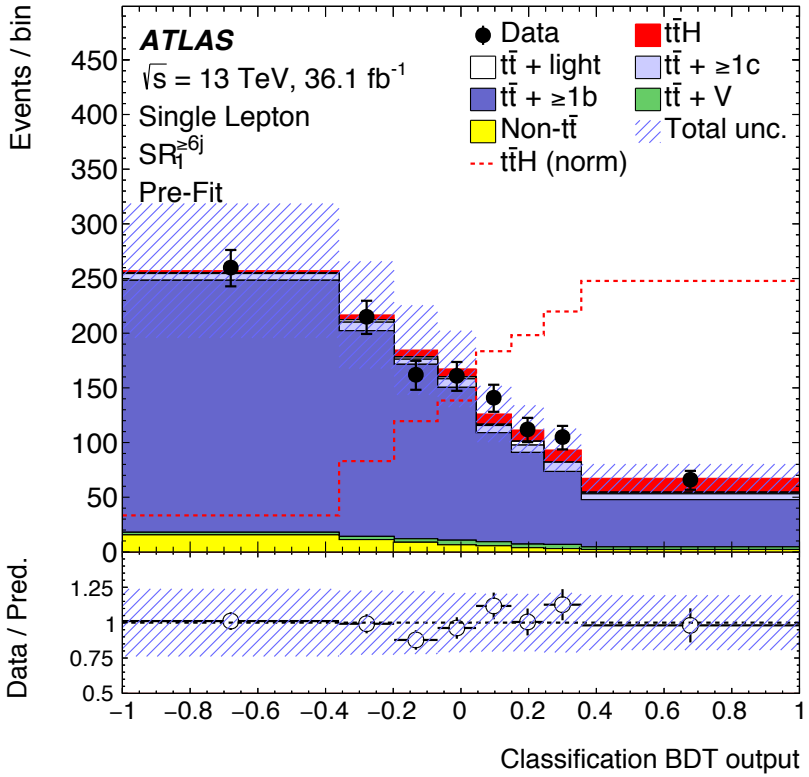
examples from HIGG-2017-93, using output from TRExFitter

- **TRExFitter** provides data/MC plots and yields per region (channel), summary plots, background composition, S/B, etc.
  - Can customize appearance for publication-quality figures



Pre-fit plots and binning

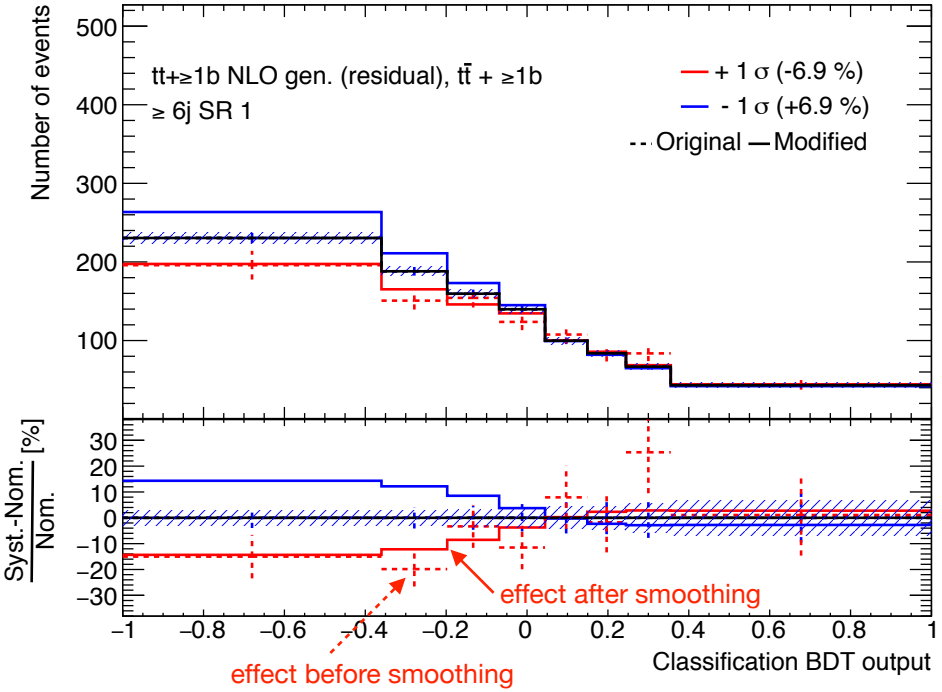
example from HIGG-2017-93, binning with TransfoD



- **One plot** like this generated **per analysis region (channel)**
- **Total uncertainty** of all sources evaluated and visualized
- **Algorithms** to automatically obtain **suitable binning**
  - Especially useful for MVA output distributions
  - Can of course also specify bins by hand

Systematic plots and smoothing

- **Visualize** the **effect** of all **systematic variations**
  - Per Region (channel), per sample, per variation
  - Important to validate the physics
- Can apply **smoothing algorithms** to systematic variations
  - **Mitigate** artificial effects from **statistical fluctuations**
  - Can help avoid strange pulls and constraints
  - Note: can also apply smoothing to samples
- In practice, there are **thousands of these distributions**
  - Can flag some problematic ones automatically, but generally there is a lot of work for the user to go through one by one



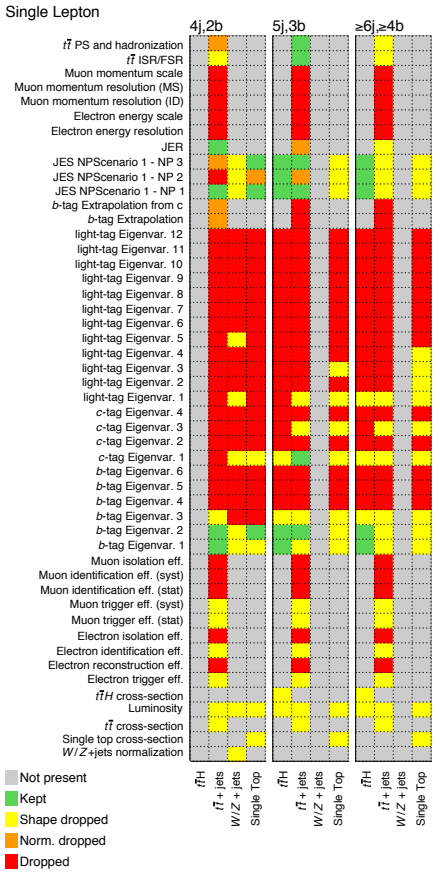
Workspace and pruning

- The **workspace** is used by HistFactory, contains all information needed to run fits

```
<!DOCTYPE Channel  SYSTEM 'HistFactorySchema.dtd'>

  <Channel Name="ljets_HThad_ge6jge4b" InputFile="" >

    <Data HistoName="ljets_HThad_ge6jge4b_Data_regBin" InputFile="ttH2015_ljets/Histograms/ttH2015_ljets_ljets_HThad_ge6jge4b_histos.root" HistoPath=""  />

    <StatErrorConfig RelErrorThreshold="0.001" ConstraintType="Poisson" />

    <Sample Name="ttH"  HistoPath=""  HistoName="ljets_HThad_ge6jge4b_ttH_regBin"  InputFile="ttH2015_ljets/Histograms/ttH2015_ljets_ljets_HThad_ge6jge4b_histos.root"  NormalizeByTheory="True" >
      <OverallSys Name="Lumi"  High="1.05"  Low="0.95"   />
      <OverallSys Name="ttHXsec"  High="1.1"  Low="0.9"   />
      <OverallSys Name="BTag_B_NP1"  High="0.790622"  Low="1.20938"   />
      <OverallSys Name="BTag_B_NP2"  High="1.03738"  Low="0.962617"   />
      <OverallSys Name="BTag_B_NP3"  High="0.979003"  Low="1.021"   />
      <OverallSys Name="BTag_C_NP1"  High="0.974812"  Low="1.02519"   />
      <OverallSys Name="BTag_Light_NP1"  High="0.98677"  Low="1.01323"   />
      <OverallSys Name="JES_Scenario1_NP1"  High="1.10107"  Low="0.898927"   />
      <OverallSys Name="JES_Scenario1_NP2"  High="1.01489"  Low="0.985106"   />
      <OverallSys Name="JES_Scenario1_NP3"  High="1.04096"  Low="0.959043"   />
      <NormFactor Name="SigXsecOverSM"  Val="1"  High="100"  Low="-100"  Const="False"   />
      <NormFactor Name="muLumi"  Val="1"  High="100"  Low="0"  Const="False"   />
```

xml files containing fit model

- At this stage, **pruning** also takes place
  - **Fit model** is **simplified** by removing small systematic variations
  - Validity of pruning needs to be verified by hand!



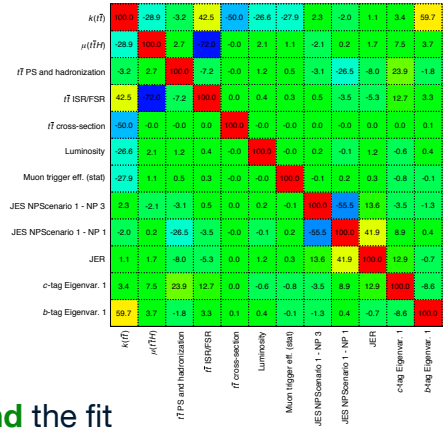
Fit results and post-fit plots

- Core framework task: **run a profile likelihood fit**
  - Many **configuration possibilities:**
    - Data or Asimov (pseudo-) data
    - Including a signal or background-only
    - Which regions to include

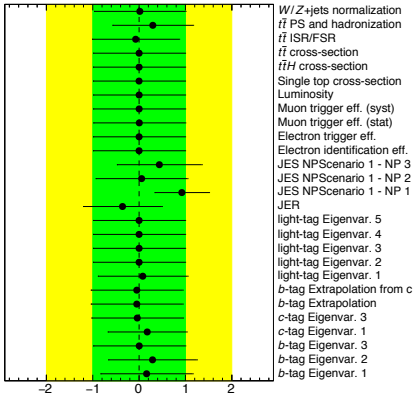
- Many **plots and files** generated to **document and understand** the fit
  - **Best-fit** values of all nuisance parameters and **associated uncertainties**
  - **Correlations** of **fit parameters**

- Can produce **post-fit plots**
  - Fit results project onto all regions defined
  - Can even validate post-fit modelling in regions not used in the fit itself

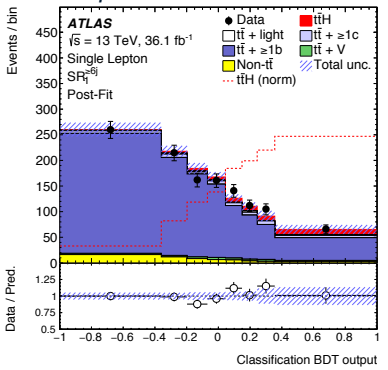
correlation matrix



nuisance parameter results



post-fit plot, example from HIGG-2017-93



Additional features (1)

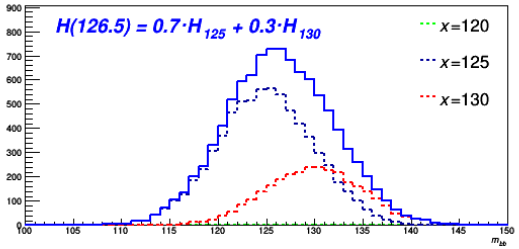
- **Lots of features** implemented **beyond a simple fit:**
  - **CL_S limits**
  - **Significance** calculation
  - **Nuisance parameters ranking** by *impact*
    - Or combined impact of nuisance parameter groups
  - **Combination** and **comparison of different fits**
  - **Toys to evaluate effect of statistical fluctuations in templates defining systematic uncertainties**
  - **Template fitting / morphing**

- Many **options** available **to understand fit behaviour better:**
  - **Exclude** nuisance parameters or **fix** them to specific values
  - **Correlate** or **de-correlate** nuisance parameters
  - Create **custom Asimov** datasets and fit them

grouped impact of systematic uncertainties, used in HIGG-2017-93

Uncertainty source	$\Delta\mu$	
$t\bar{t}+\geq 1b$ modeling	+0.46	−0.46
Background-model stat. unc.	+0.29	−0.31
b-tagging efficiency and mis-tag rates	+0.16	−0.16
Jet energy scale and resolution	+0.14	−0.14
$t\bar{t}H$ modeling	+0.22	−0.05
$t\bar{t}+\geq 1c$ modeling	+0.09	−0.11
JVT, pileup modeling	+0.03	−0.05
Other background modeling	+0.08	−0.08
$t\bar{t}+$ light modeling	+0.06	−0.03
Luminosity	+0.03	−0.02
Light lepton (e,μ) id., isolation, trigger	+0.03	−0.04
Total systematic uncertainty	+0.57	−0.54
$t\bar{t}+\geq 1b$ normalization	+0.09	−0.10
$t\bar{t}+\geq 1c$ normalization	+0.02	−0.03
Intrinsic statistical uncertainty	+0.21	−0.20
Total statistical uncertainty	+0.29	−0.29
Total uncertainty	+0.64	−0.61

signal template example



- **Statistical analysis frameworks:**
  - Declare a fit model in a human-readable way and provide relevant inputs
  - Framework builds HistFactory workspace and everything needed for fits and validation
  - Implement many tools to help analyses study fits in detail