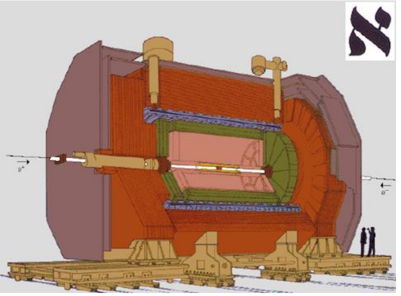
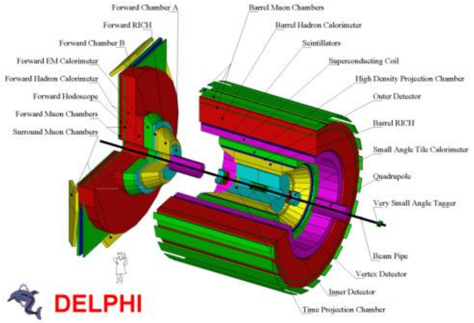
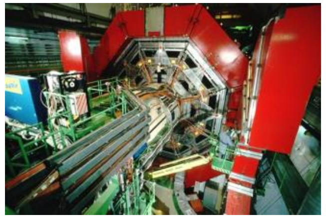
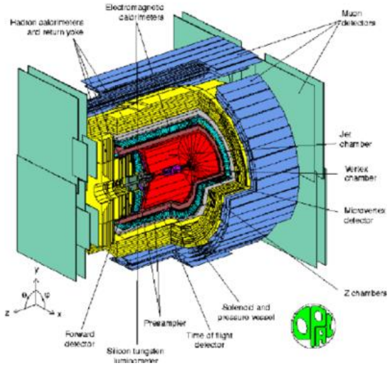
Logistics of Making Combined Results









Thomas R. Junk

Fermilab

CMS LPC EFT Workshop #2

Notre Dame, April 25, 2024

Examples I am Most Familiar With

LEP2: Searches for SM and MSSM Higgs bosons

- I prepared OPAL's results for combination, and did the MSSM Higgs limit calculations for OPAL and for the LEP combination (ALEPH + DELPHI + L3 + OPAL)
- I created tools for preparing inputs and integrated tools from other contributions – Many thanks to Alex Read for histogram morphing and for developing CLs!

LEP1: Precision Electroweak Measurements

- A very different sort of combination work from the searches
- While I didn't perform the calculations, I did review the procedure -- it's not that complicated. (BLUE)

Tevatron:

- Single Top search and discovery (2009). Many thanks to Kevin Lannon for doing the actual work!
- Tool development -- mclimit_csm.C originally developed for setting limits and discovering single top.

- SM Higgs boson search @ the Tevatron – input collection, packaging, calculation of CDF's limits' p-values, and best-fit signal strength (now customarily called μ; we called it R). Combination with D0

DUNE:

- I wrote an oscillation parameter fitter, though there are several competing ones. Big issues are understanding systematic uncertainties, not in fitter development. I am the liaison between the computing consortium and the physics groups, and one of DUNE's software managers.

Setting the Stage: Basic Assumptions (and when they are false)

Most Collider and fixed-target analyses have the following features:

1) Data are in the form of counts of events drawn from Poisson distributions.

2) Each counted interaction may have many measured quantities, or possibly none at all.
   Statisticians call these "marked Poisson" data.

3) Model predictions add incoherently:
   Total Predicted Rate = Signal + Background

   Statisticians call these "mixture models".

Setting the Stage: Basic Assumptions (and when they are false)

Most Collider and fixed-target analyses have the following features:

1) Data are in the form of counts of events drawn from Poisson distributions.

Not true when data events are weighted (hardly anyone does this)

Monte Carlo events are weighted all the time, however. Doesn't break the assumption though.

Precision electroweak combinations did not collect data in raw Poisson form but rather finished parameter estimates. Parameter estimates are not "data" but rather are functions of the data and the models used to interpret them.

Setting the Stage: Basic Assumptions (and when they are false)

2) Each counted interaction may have many measured quantities, or possibly none at all. Statisticians call these "marked Poisson" data.

In the case of weighted data events, one can treat those weights as reconstructed variables and simply bin the data in the reconstructed variables and still count each event singly.

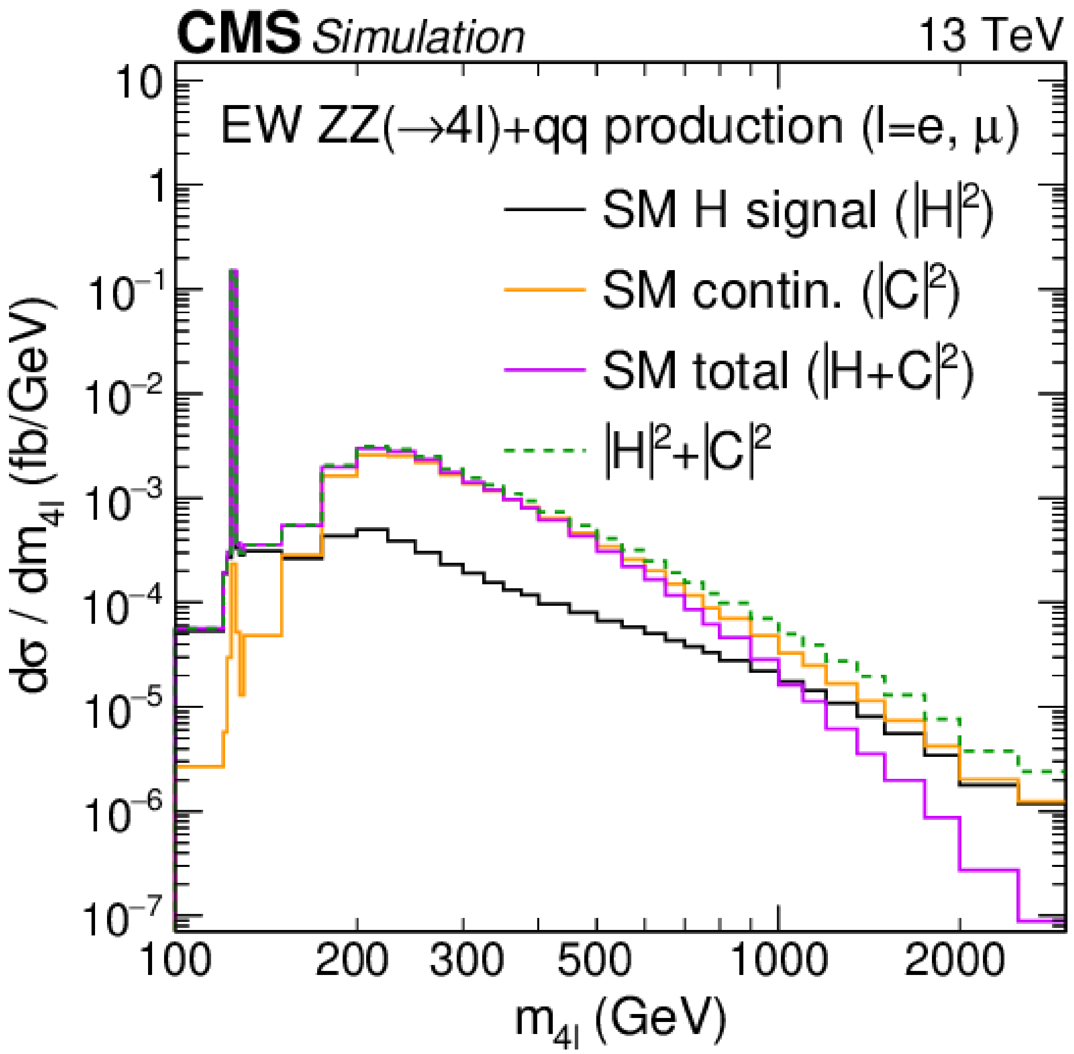
Setting the Stage: Basic Assumptions (and when they are false)

3) Model predictions add incoherently:
Total Predicted Rate = Signal + Background

Quantum Mechanics tells us total rates are quadratic functions of coupling constants. See Kelci Morhman's talk on Monday.

Sometimes, adding a new process reduces the total predicted event rate

Example: Off-shell Higgs Boson production interferes destructively with the background.



CMS Collab., *Nature Phys.* 18 (2022) 11, 1329-1334 e-Print: 2202.06923 [hep-ex]

Setting the Stage: Basic Assumptions (and when they are false)

3) Model predictions add incoherently:
Total Predicted Rate = Signal + Background

Decay Branching Fractions add to 1.0.

Exotic particles that are hard to detect may reduce the count of visible decays.

Example: Top quark decays to charged Higgs bosons and b quarks. If the charged Higgs decays differently compared to a W boson, it can reduce event yields in targeted selections.

CDF Collab., *Phys.Rev.Lett.* 96 (2006) 042003 e-Print: hep-ex/0510065 [hep-ex]

Setting the Stage: Basic Assumptions (and when they are false)

3') Model predictions add incoherently and scale with a rate parameter μ

Total Predicted Rate = μ*Signal + Background

Sometimes, to get a larger signal rate out of a model, we must increase coupling strengths.

Example: SUSY bbH → bbbb at CDF. Coupling of H to bb is proportional to tanβ. So the production rate scaled with $\tan^2\beta$, but so did the decay width.

Signal got larger, but also wider. Looking for a peak on a smooth background gets harder when the peak is wide, and even ends up looking just like the background shape.

CDF Collab., *Phys.Rev.D* 85 (2012) 032005 e-Print: 1106.4782 [hep-ex]

Cut-And-Count Searches for New Physics

Pre-1990's, and even well into the LEP2 era, most searches were done this way.

Some still are!

"Recent" example: DONUT's observation of the Tau Neutrino:
Phys.Lett.B 504 (2001) 218-224 e-Print: hep-ex/0012035 [hep-ex]

$n_{obs} = 4$
$b = 0.34 \pm 0.05$
$p = 4 \times 10^{-4}$ (not quite 5σ, but more than 4σ)

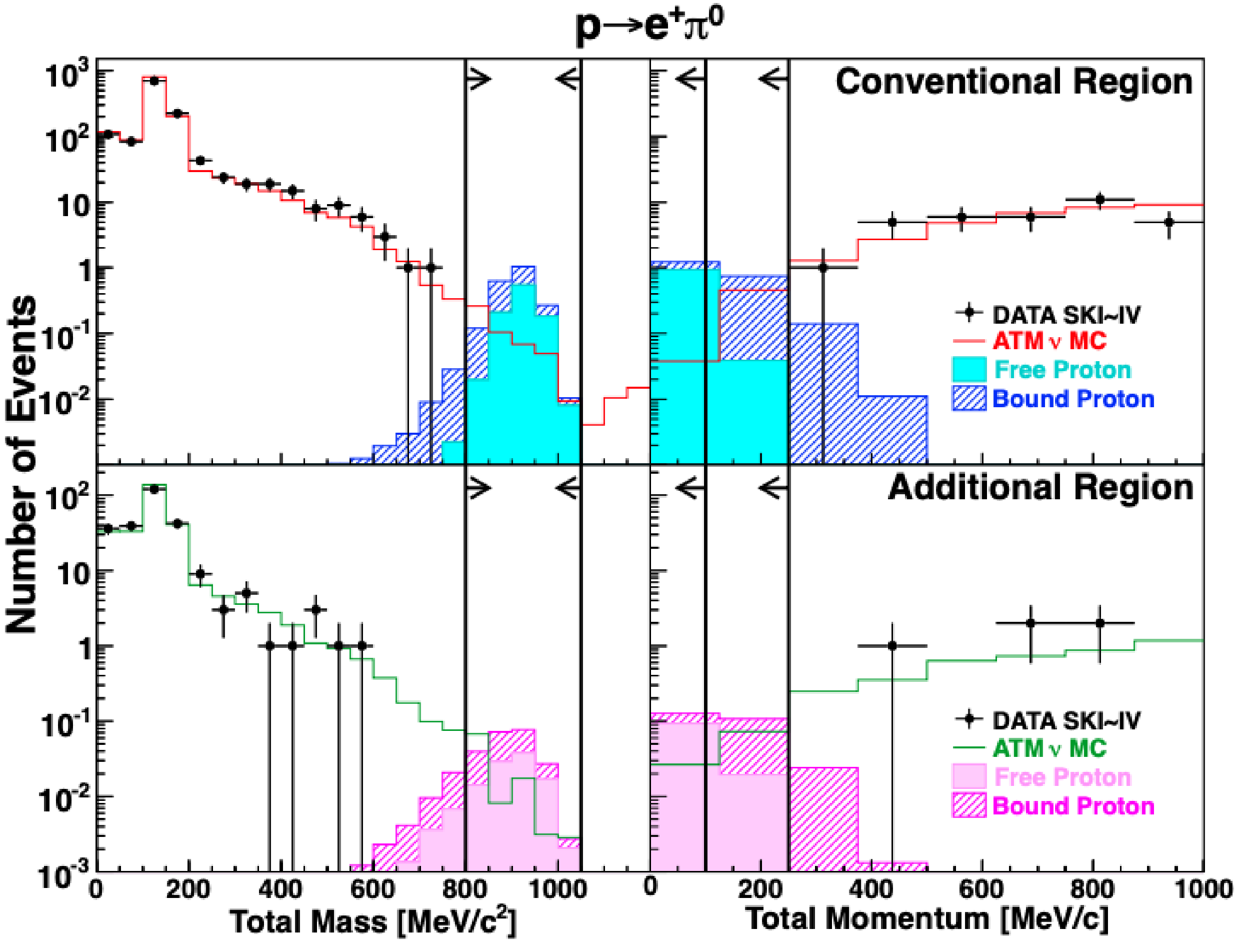
It always seemed as if at LEP2, cut-and-count searches for new physics, after optimizing cuts, always seemed to have background estimates of order 1 event.

This is not entirely coincidental – zero-background searches have expected cross-section limits that scale with integrated luminosity (or "exposure" if you work in those units).

"Almost" cut and count – Eight subsamples in a Proton Decay Search by Super-K

Each one is a cut-and-count analysis, but they are combined by multiplying the Poisson likelihoods.

Background prediction falls very steeply and signal falls off – little to be gained by digging into the background, and lose 1/exposure scaling.



Super-Kamiokande Collab., *Phys.Rev.D* 102 (2020) 11, 112011 e-Print: 2010.16098 [hep-ex]

Optimizing Cut-And-Count Analyses

Optimal operating points often differ for discovery and exclusion due to different criteria.

Exclusion at 95% CL requires an expectation of 3 events of signal, even if there is no expected background, and no events observed.

Discovery requires significant excess over background – best when background is small.

$b = 3\times10^{-7}$ and $n_{obs} = 1$ is a discovery. (Ω^- was discovered with one bubble-chamber event).

You can discover with a bigger background expectation with more observed events, but still the premium is on low background.

See G. Punzi, *eConf* C030908 (2003) MODT002 Contribution to:
PHYSTAT (2003): Statistical Problems in Particle Physics, Astrophysics, and Cosmology, 79-83 e-Print: physics/0308063 [physics]

Optimizing Cut-And-Count Analyses

In general, as you collect more integrated luminosity (or exposure), the optimal cuts get *tighter*.

Optimizing for a combination is different from optimizing an analysis in isolation (how tight to make the cuts?)

Optimizing for exclusion and discovery are a bit in tension.

These issues go away when you rank events by some MVA score and report all events, not just a single number of events after some cut.

Ideally, if we could assign a local s/b ratio for each event and each possible event, we would be optimal.

Systematic uncertainties and finite MC statistics get in the way of just reporting all n-dimensional quantities of each event.

LEP2 SM Higgs Boson Searches

Increasing sophistication in analyses

- Multiple search channels: ZH→llbb, jjbb, Pt(miss)bb, and H→tautau channels.
- Multivariate discriminants were getting popular – neural networks, BDTs, likelihood functions ("Naive Bayes")
- Good ol' m_{jj} and m_{tautau} (kinematically constrained!) were used as discriminants. Often the MVA variables were on a second axis.
- Large teams (for the time) were assembled to do the work.
- Analyzers wanted all of their hard work to propagate to improved sensitivity of the expected limits and expected p values.

Cut and Count wasn't good enough anymore!

Analyses still had cuts because the vast majority of triggers at a collider detector are not relevant for a particular search and effort spent trying to model irrelevant data is not well spent.

Preslection + MVA → statistical analysis was the workflow.

ALEPH, DELPHI, L3 and OPAL Collaboratons, *Phys.Lett.B* 565 (2003) 61-75 e-Print: hep-ex/0306033 [hep-ex]
OPAL Collab., *Eur.Phys.J.C* 26 (2003) 479-503 e-Print: hep-ex/0209078 [hep-ex]

LEP2 Higgs Boson Search Workflow Requirements

We wanted to be able to

- compute L(data|s+b) and L(data|b) for calculation of the likelihood ratio test statistic and compute other things like Bayesian posteriors. Also accommodate other statistical procedures.
- draw pseudoexperiments – needed for calculation of p values.
- include the effect of systematic uncertainties
- Ensure we were all testing the same model (shared cross section, branching ratio, background models, and systematic uncertainties).

And do so reproducibly

- Combiners must first reproduce the individual experiment's results using exchanged artifacts
- Each collaboration provided at least one combiner – methods varied, and consistency was investigated
  - "technical spread" i.e. bugs
  - method-induced spread – change the test statistic or use Bayes vs. CLs – some variation is expected.

LEP2 Higgs Boson Search Workflow Requirements

And do so quickly

- Collaborations usually wanted to present combined results at the same conferences as the results themselves
- Need to freeze analyses and exchange inputs long enough in advance to
  - inspect and debug them
  - run combination
  - write combination physics note
  - go through experiment approval procedures
- Approval was usually quick, as experiments already reviewed their results, and it wasn't up to the combination group to second-guess the experiment results. But disagreements did arise and speculation about the quality of the input analyses always happened. In the end, we just combined things.

LEP2 Higgs Boson Combination Workflow

Solution chosen: exchange binned data. Combiners don't have to choose binning or apply smoothing, improving reproducibility. No choices made to adjust data or model by combiners!

No unbinned results – just bin finer if it was needed. Empty data bins are not a problem, empty MC bins were.

In each bin, quote: signal, background, and observed data counts.

Systematic uncertainties were listed by named source.

Systematics with the same name were assumed to be 100% correlated. Others, 0% correlated.

If you needed partial correlation (e.g. a partially shared MC sample or data control sample), you could always decompose it into correlated and uncorrelated parts.

Inputs were exchanged as standalone FORTRAN 77 functions. Linked with CERNLIB (including HBOOK). Auxiliary files (data, systematics, text, .rz) were provided along with the FORTRAN files in tarballs and exchanged in a shared, private space in AFS.

Inputs: m_h, m_A, m_H, cross sections and branching ratios. (accommodates SUSY and more general 2HDMs)
Outputs: binned search results and signal and background predictions.
Histograms were "flattened" – 2D and more were represented as 1D arrays.

Template Morphing: Vertical and Horizontal

Especially needed for the LEP SM case because MC samples were generated at a fixed set of m_H values, typically spaced at least 1 GeV apart, but the resolution after the beam constraint was much finer (of order a few hundred MeV) Cross sections also were very strong functions of m_H due to the sharp kinematic edge of the maximum energy of the $e+e^-$ collisions.

An example analysis that uses $m_{H,reco}$ as a discriminant variable ("bump hunt") therefore could only be run on a coarse grid of m_H values unless an appropriate template morphing method was used.

Horizontal Morphing: See Alex Read, Linear Interpolation of Histograms, *Nucl.Instrum.Meth.A* 425 (1999) 357-360.

Also known as "optimal transport"

Vertical morphing – just interpolate within each bin. Useful for NN and other MVA variables where horizontal morphing would produce spurious intermediate peaks in multimodal (signal, background) distributions.

Template morphing was done in experiment-provided code, not done by combiners.

Demo: te(), te(1), te2() in https://github.com/tomjunk/mclimit/blob/main/tesyst.C

Template Morphing

For a recent review, see Lydia Brenner's presentation at the
BANFF Phystat meeting in 2023:
https://www.birs.ca/workshops/2023/23w5096/files/Brenner/Brenner-TemplateMorphingBanff.pdf

The section on effective Lagrangian morphing may be of special interest to this group.

More documentation:

Balasubramanian, Brenner, Burgard, and Verkerke,
"Effective Lagrangian Morphing" https://arxiv.org/abs/2202.13612

MVA Analysis Optimization – Spaghetti Plots

ALEPH, DELPHI, L3 and OPAL Collaboratons,
Phys.Lett.B 565 (2003) 61-75
e-Print: hep-ex/0306033 [hep-ex]

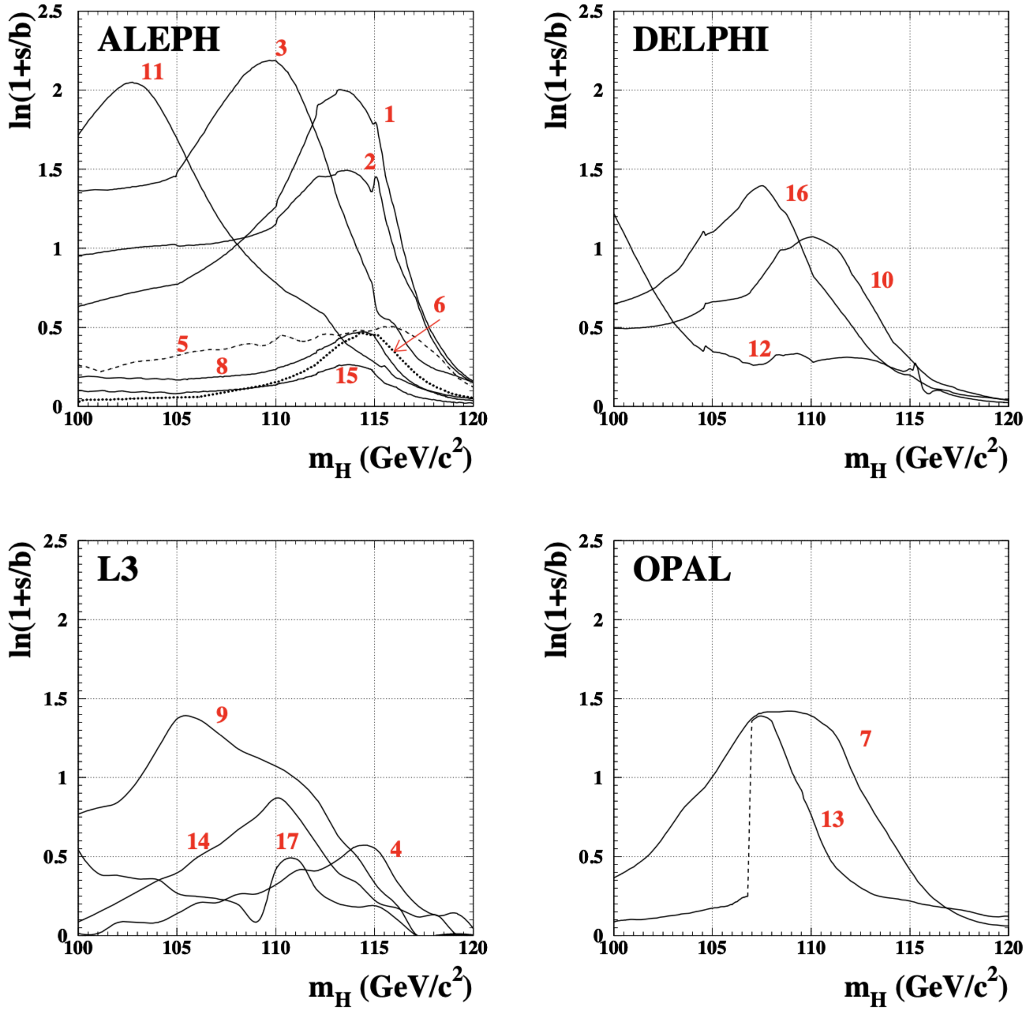


Figure 3: Evolution of the event weight $\ln(1 + s/b)$ with test mass m_{H} for the events which have the largest contributions to $-2\ln Q$ at $m_{\mathrm{H}} = 115$ GeV/c^2. The labels correspond to the candidate numbers in the first column of Table 2. The sudden increase in the weight of the OPAL missing energy candidate labelled "13" at $m_{\mathrm{H}} = 107$ GeV/c^2 is due to switching from the low-mass to high-mass optimization of the search at that mass. A similar increase is observed in the case of the L3 four-jet candidate labelled "17" which is due to the test mass dependent attribution of the jets to the Z and Higgs bosons.

LEP2 Higgs Boson Searches: MSSM

There was a lot of re-use of the SM Higgs boson search data as inputs to MSSM Higgs boson searches.

The ZH final state was the same. (Zh and ZH had the same kinematics, just different production rates and Higgs boson masses).

Because MVAs could depend on mH, we could not simply add signals to the same background histogram (more on this later). Nearly all models (all that I could think of) had mH out of kinematic reach, and we only assumed one CP-even Higgs boson

Extra channels: Z->hA (4b and bbtautau)

Shared model database: Arnulf Quadt ran a large number of predictions with FeynHiggs and made an HBOOK ntuple out of it. Scans over mA and tanbeta, but there were several scenarios considered.

http://www.feynhiggs.de/
arXiv:1811.09073 [hep-ph], arXiv:1706.00346 [hep-ph], arXiv:1608.01880 [hep-ph], arXiv:1312.4937 [hep-ph], hep-ph/0611326, hep-ph/0212020, hep-ph/9812472, hep-ph/9812320.

MSSM Combination: ALEPH, DELPHI, L3, OPAL: *Eur.Phys.J.C* 47 (2006) 547-587 https://arxiv.org/abs/hep-ex/0602042
OPAL Collab: *Eur.Phys.J.C* 37 (2004) 49-78 https://arxiv.org/abs/hep-ex/0406057

A New Tool for Use at Hadron Colliders

I wrote a new tool in C++ using ROOT to calculate p-values, CLs, and Bayesian posteriors, called mclimit_csm.C. (okay, it was new in 2006).

Better handling of systematic uncertainties – now nuisance parameters were fit for in each calculation of the likelihood – the "profile" likelihood.

Which parameters you fit for is up to the user – not every parameter needed fitting.

If you were fitting for MC stats in each bin, it had a Barlow-Beeston (*Comput.Phys.Commun.* 77 (1993) 219-228) calculator.

Horizontal and vertical template morphing were provided.

Parameters were fluctuated in pseudoexperiments. One ought to fluctuate the constraints and fix the parameters, but in practice it didn't make much difference. At LEP, we fluctuated but did not fit. At the Tevatron, signals were smaller than the a priori uncertainties on the background, so fitting was mandatory.

Bayesian marginalization tools also provided using Metropolis-Hastings MCMC with a proposal function that is optimized using the input s and b predictions in a pre-scan over parameter space.

Tool used first in single top analyses. Very similar to WH$\rightarrow$lvbb search at the Tevatron.

Compounding Shape Morphing

Old Strategy: Interpolate each shape systematic from the central value, and average the resulting interpolated histograms

- Commutative, at least
- Biased towards central shape: Add in a shape uncertainty which is small, and it averages in like the big shape uncertainties.

- Want: SAT-style “analogy” interpolation

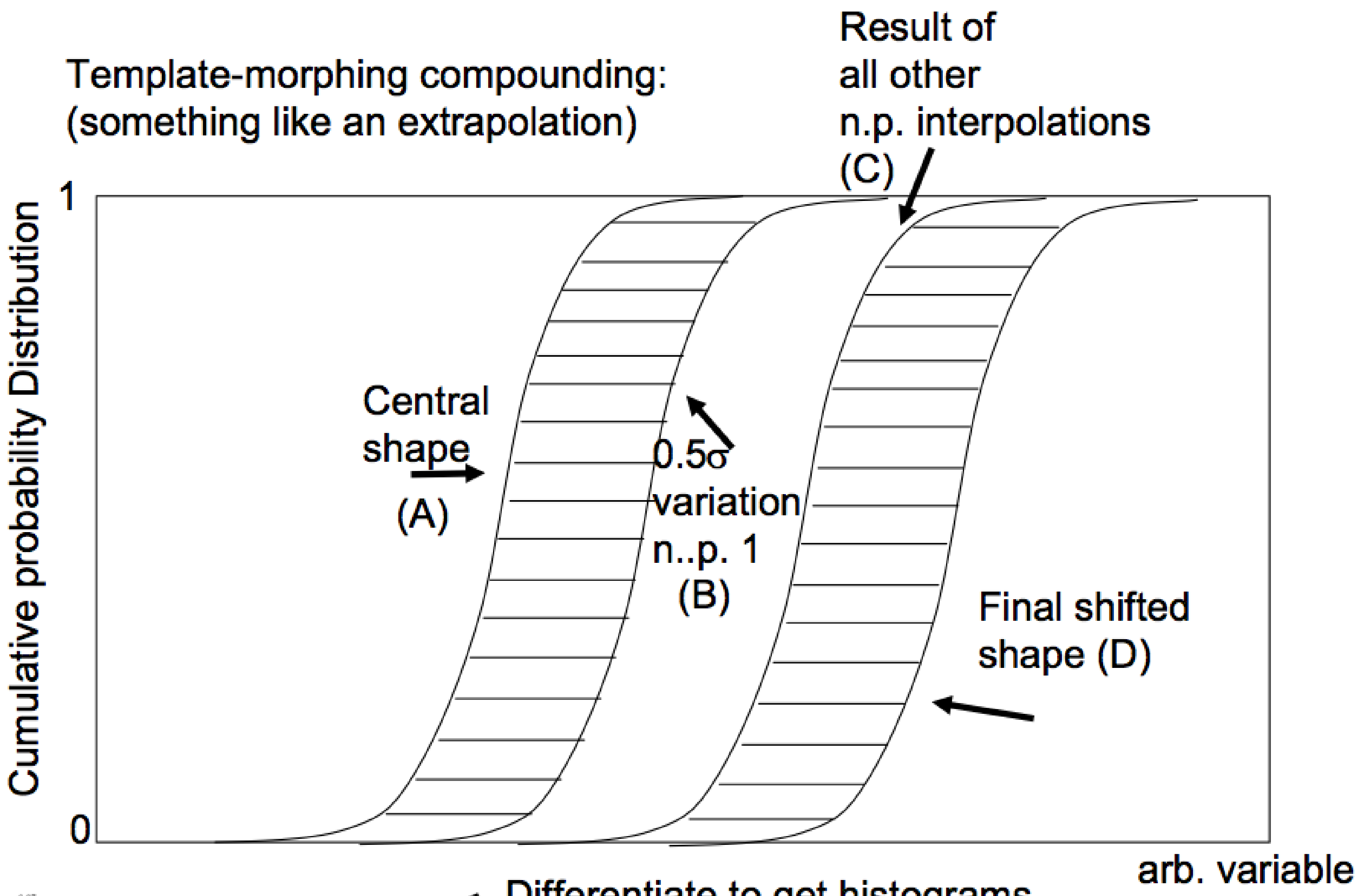
A:B as C:D, where

A = central value
B = systematically distorted template
C = result of previous distortions from other nuisance parameters
D = result of all distortions put together.

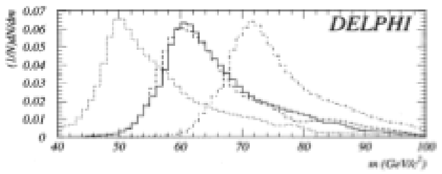
With one free parameter – how much to vary the nuisance parameter under consideration.

Shape Systematic Compounding Procedure in mclimit_csm.C

Template-morphing compounding:
(something like an extrapolation)



Demo: *tesyst()*



Differentiate to get histograms

Commutative

Commutative! Extrapolates!

Optimizing MVA Analyses

This is usually done with a *loss function*.

Often it is something simple like the sum of squares of classification or regression errors. Others are possible, but to be used in training, it is desirable to be able to differentiate it with respect to an MVA parameter (like a NN weight).

We almost never care about optimizing this particular loss function.

We care about discovering new particles and interactions, and measuring their properties.

Desired figures of merit: median expected p-value, median expected upper limit, and expected uncertainty on measured parameters (like cross sections and branching fractions).

Optimization is therefore not impossible, simply difficult – try many methods and pick the ones with the best metrics.

MVAs reduce the tension between exclusion and discovery optimization. If there were no systematic uncertainties and we could bin our data finely in each variable, we'd be done with the optimization.

MVAs are great dimensionality-reducing tools.

Tevatron: Single Top Search and Discovery

Different logistics!
Large teams worked on single top physics. CDF & D0, but no CDF+D0 combination!
We expected to be able to discover single top separately. (later, *s*-channel single top was observed in combination)
Lots to do: B-tag development, calibration, lepton ID and energy calibration,
MC generation, ntuple making, analysis optimization, statistical interpretation

Everyone wanted to discover single top with their favorite MVA methods!

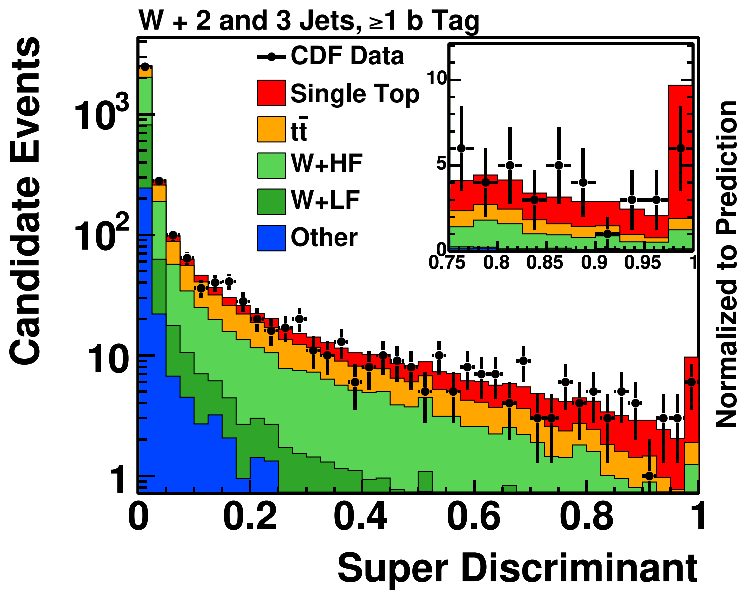
Old Likelihood style
Neural Networks
Matrix Element methods
Boosted Decision Trees

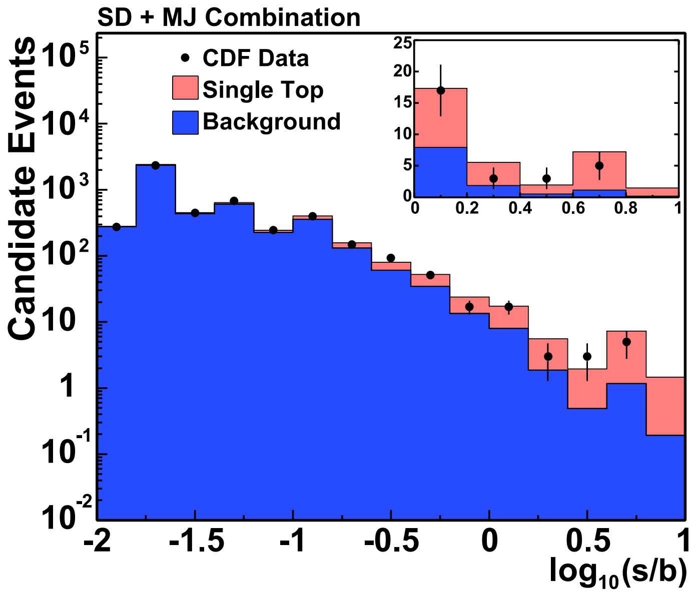
We couldn't publish each result separately – we might have two discoveries and two near misses with the same data.

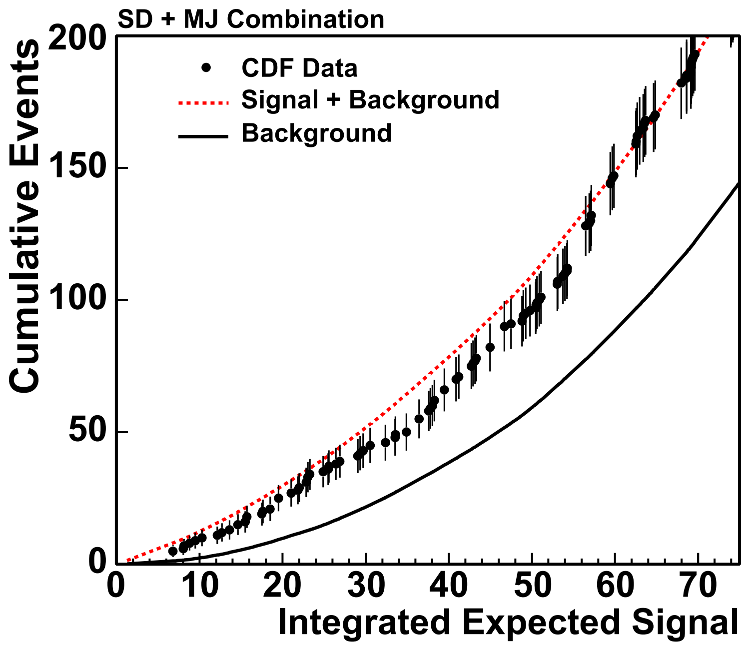
Instead, we computed a Super Discriminant using NEAT (Neuro-Evolution of Augmenting Topologies), which assigned a single score to each event based on the other MVA scores.
Stanley and Miikkulainen, *Evolutionary Computation*, 10(2):99-127, 2002.
https://nn.cs.utexas.edu/?stanley:ec02

Kevin Lannon is our hero for implementing all this!

MVAs and Answering Questions About Cut and Count







CDF Collab., *Phys.Rev.D* 82 (2010) 112005 e-Print: 1004.1181 [hep-ex]

CDF and D0 Single Top References

D0 Evidence PRD: *Phys.Rev.D* 78 (2008) 012005 e-Print: 0803.0739 [hep-ex]

D0 Observation Letter: *Phys.Rev.Lett.* 103 (2009) 092001 e-Print: 0903.0850 [hep-ex]
Follow-Up PRD: *Phys.Rev.D* 84 (2011) 112001 e-Print: 1108.3091 [hep-ex]

CDF Observation Letter: *Phys.Rev.Lett.* 103 (2009) 092002 e-Print: 0903.0885 [hep-ex]
Follow-Up PRD: *Phys.Rev.D* 82 (2010) 112005 e-Print: 1004.1181 [hep-ex]

Observation of *s*-Channel Single top: CDF and D0 Combination:
CDF and D0 Collaborations, *Phys.Rev.Lett.* 112 (2014) 231803

Tevatron Higgs Searches

Multiple possible production methods and decay modes meant that many searches were needed.
Discovery would be in combination, even within a collaboration.

lvbb, llbb, METbb, lvtautau, lltautau, METtautau,
WW->lvlv, H→ZZ→ll, and H→gammagamma

Inputs exchanged as before – binned, with alternate shapes and rate systematic uncertainties listed by named input source.

Pre-agreed mass grid. No interpolation on m_H – MVAs optimized separately at each m_H. No spaghetti plots!

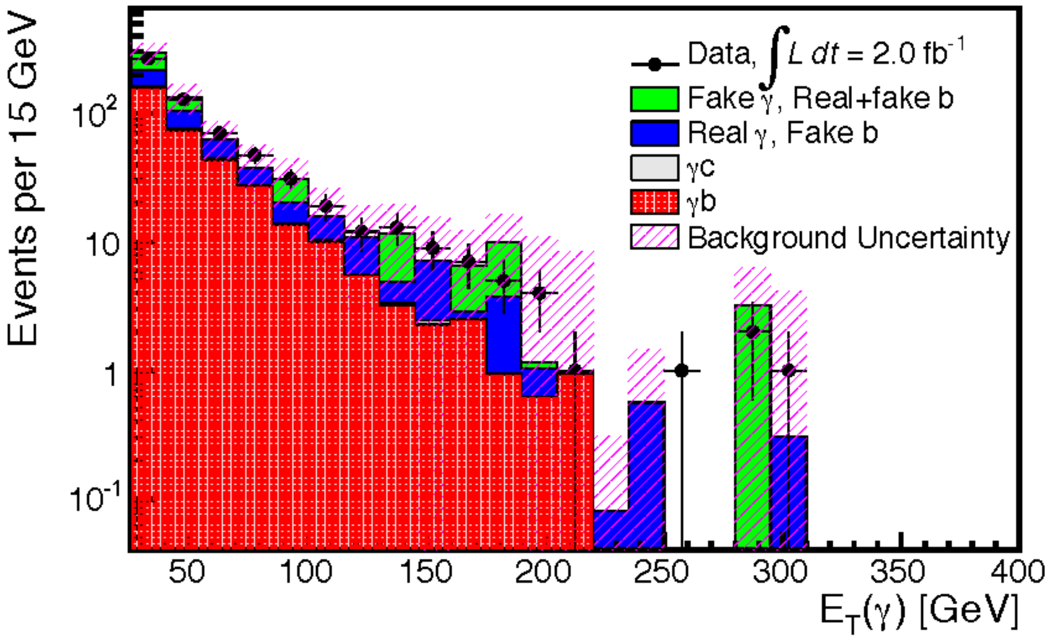
Cross sections and branching ratios for the signal and main backgrounds agreed upon within an experiment and between experiments beforehand.

Bigger job – many more channels and bins and systematics.

Also: signals and backgrounds are now stacked templates – not just a total signal and total background –
Needed to get systematic uncertainties right! Some uncertainties affect some backgrounds and not others.

Validation of inputs was a big task:
Plots and tables need to be made by the statistical analysis tool to check for input preparation mistakes.

MC Statistics and "Broken" Bins



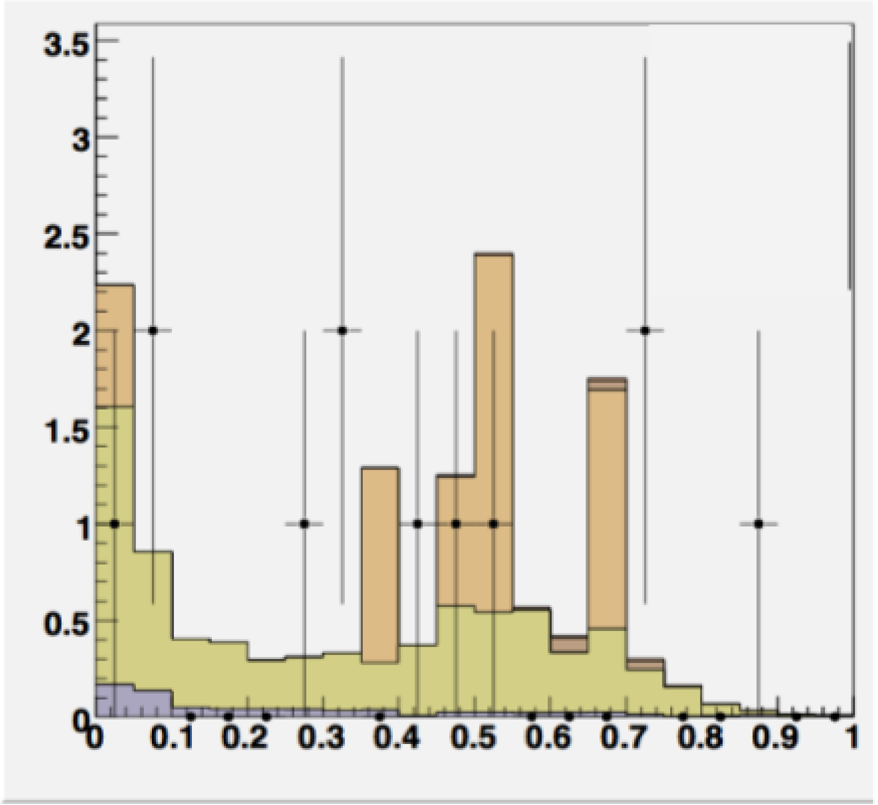
NDOF=?

• Automated tools cannot tell if the background expectation is really zero or just a downward MC fluctuation.
• Real background estimations are sums of predictions with very different weights in each MC event (or data event)
• Rebinning or just collecting the last few bins together often helps.

• Advice: Make your own visible underflow and overflow bins do not rely on ROOT's underflow/overflow bins -- they are usually not plotted. Tools may ignore u/o bins

"Partially" Broken Bins? How Can we Tell the Bins are Broken?

An Extreme Example (names removed)



Questions: What's the shape we are trying to estimate?
What is the uncertainty on that shape?

There many not be enough information in this histogram to determine shape.

One bin may be right answer.

Orange contribution was estimated from a data sideband – hard just to run some more MC to fix the problem!

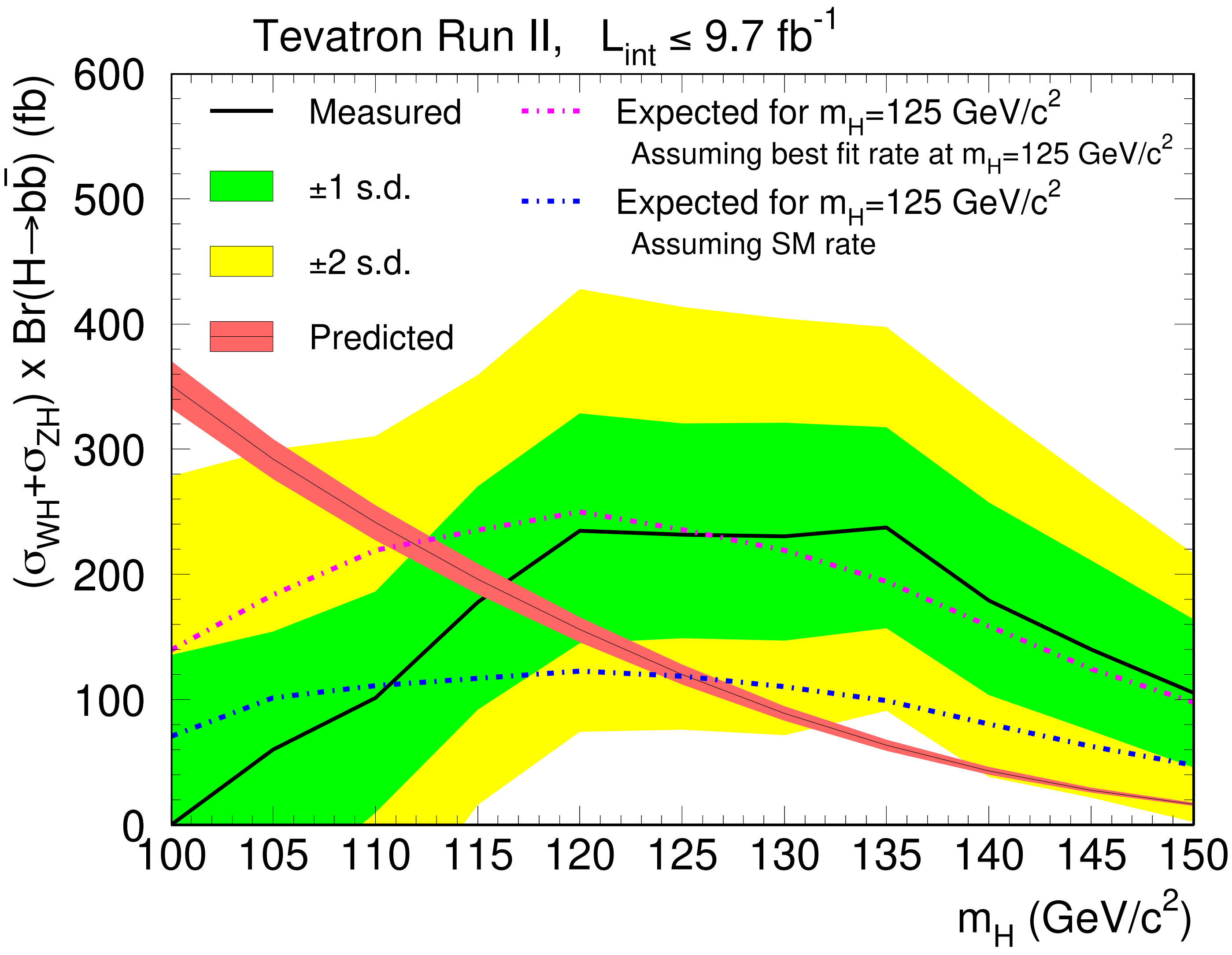
You could try smoothing this or collapsing it into one bin.
But combiners should not make these choices. Review results and send back to the analyzers!

Signal Injection Studies

Example: Studies of the Higgs boson at the Tevatron

CDF and D0 Collaborations,
Phys.Rev.D 88 (2013) 5, 052014
e-Print: [1303.6346] [hep-ex]

"What would the search for one particle look like if another particle is truly present?"



Re-using and Recasting Results

- The final state of a sought interaction may be the same as a different one.
- The differences may be simple enough that the signal can simply be scaled to its new rate;
- Trigger, reconstruction, selection and analysis efficiencies may be strong functions of the event kinematics however.

- Example: Tevatron Higgs spin and parity tests.
  CDF and D0 Collborations, *Phys.Rev.Lett.* 114 (2015) 15, 151802
  e-Print: 1502.00967 [hep-ex]

- Can we re-use a highly-tuned MVA analysis looking for the SM Higgs boson as a search for a $J^P = 0^-$ or 2^+ Higgs-like particle?

Higgs Spin and Parity – SM Higgs Search Results Cannot be Recast – Need a New Analysis

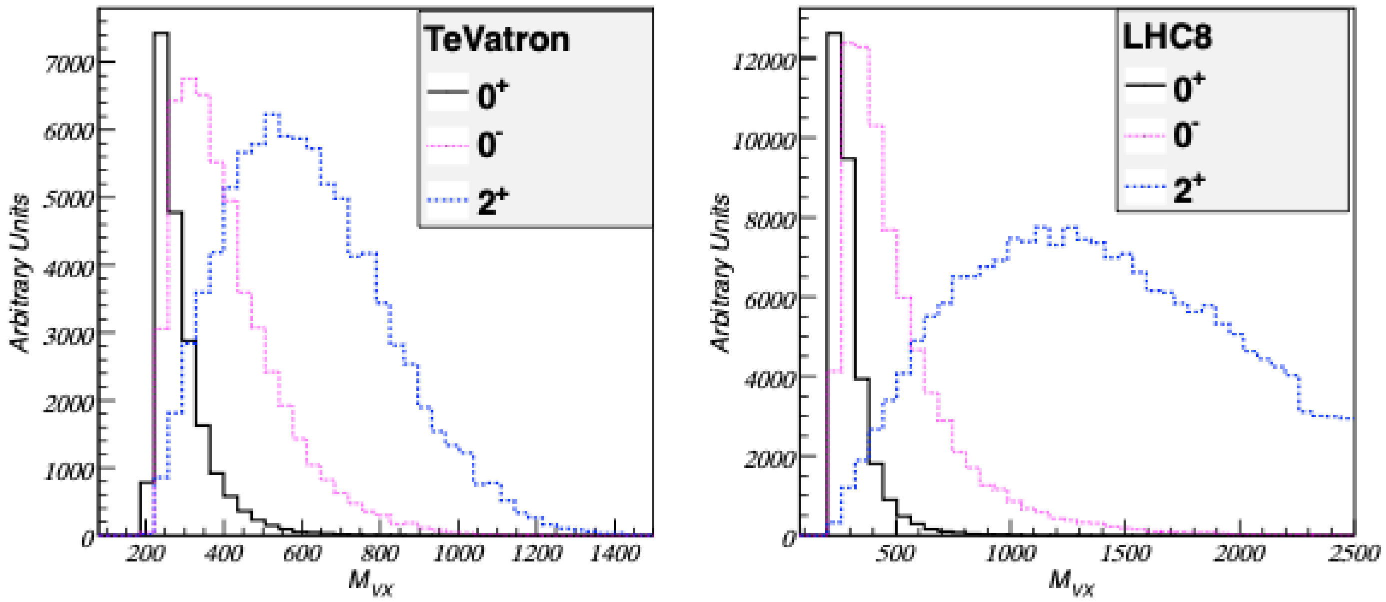


FIG. 1: *The distributions in the $Z + X$ invariant mass M_{ZX} for the 0^+ (solid black), 0^- (pink dotted) and 2^+ (blue dashed) assignments for the particle X with mass $\sim$ 125 GeV discovered by ATLAS [1] and CMS [2], calculated for the reaction $\bar{p}p \to Z + X$ at the TeVatron (left) and for the reaction $pp \to Z + X$ at the LHC at 8 TeV (right).*

Ellis, Hwang, Sanz and You, *JHEP* 11 (2012) 134, e-Print: 1208.6002 [hep-ph]

Experiment Acceptance and Analysis Cuts can Help Make Signals Look more Alike

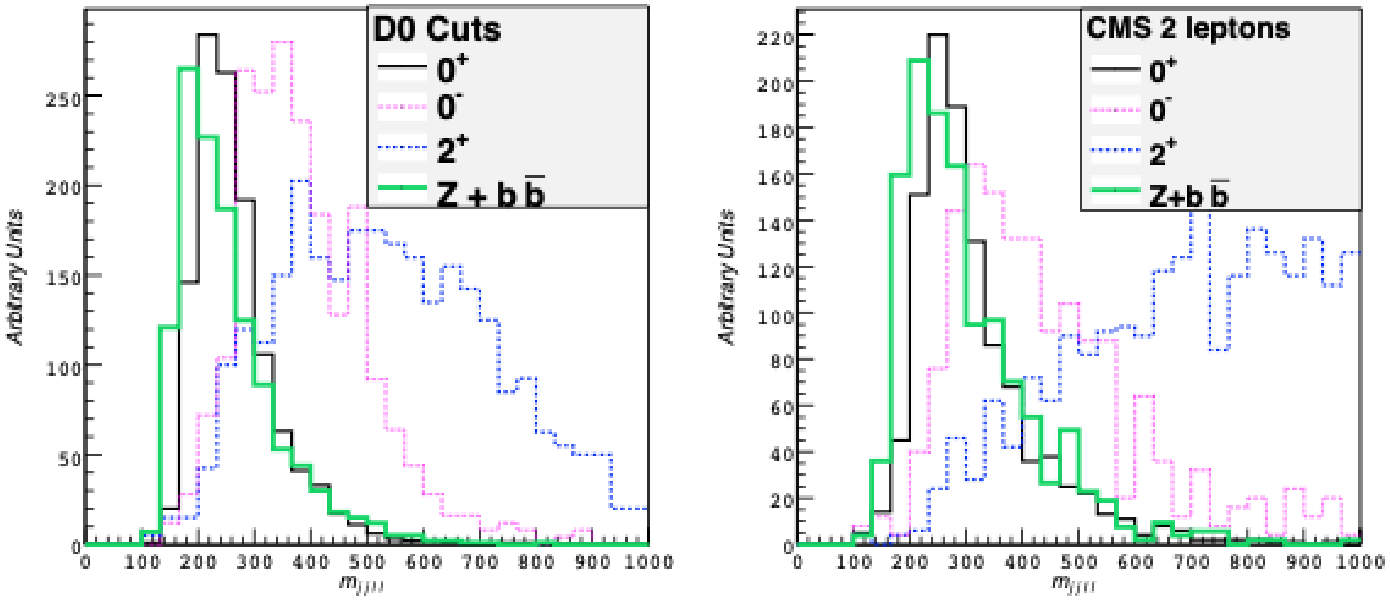


FIG. 2: *The $Z + \bar{b}b$ background invariant mass distribution (green) at the TeVatron using the D0 cuts described in the text (left panel) and the LHC at 8 TeV using the CMS cuts also described in the text (right panel) compared with the two-lepton signal distributions in the $Z + X$ invariant mass M_{ZX} for the 0^+ (solid black), 0^- (pink dotted) and 2^+ (blue dashed) assignments for the particle X with mass ~ 125 GeV.*

MVAs and Specificity

To optimize an analysis for a specific signal, one trains with an idea of exactly what that signal is

Story about training a classifier to separate pictures of tanks and trucks: training sample had a domain shift with respect to the data it was presented with.

Domain shift in the signal part can make identification efficiency less good –   cross section measurements come out too low because we are looking for something not quite what is actually there and only seeing part of it.

We symmetrize a lot of systematic uncertainties, but propagate these through the full p-value and cross section calculations.

Newer Tools: RooFit and RooWorkspaces

I am not an expert in these, but they were recommended for use in the Higgs discovery at the LHC

ATLAS and CMS Collaborations, and the LHC Higgs Combination Group,
"Procedure for the LHC Higgs boson search combination in Summer 2011"

ATL-PHYS-PUB-2011-11
CMS NOTE-2011/005

Includes the ability to provide unbinned search results.

Examples of Recast Analyses

Tools for recasting analyses:

RECAST: K. Cranmer and I. Yavin JHEP 04, 038 (2011). arXiv:1010.2506
ATLAS Analysis Preservation: K. Cranmer and L. Heinrich, J. Phys. Conf. Ser. 1085, 4, 042011 (2018)
RIVET: A. Buckley et al., Comput. Phys. Commun. 184, 2803 (2013). arXiv:1003.0694

For a discussion, see:

Junk, T., & Lyons, L. (2020). Reproducibility and Replication of Experimental Particle Physics Results. *Harvard Data Science Review*, *2*(4). https://doi.org/10.1162/99608f92.250f995b

and references therein. Online link:
https://hdsr.mitpress.mit.edu/pub/1lhu0zvn/release/4?readingCollection=c6cf45bb

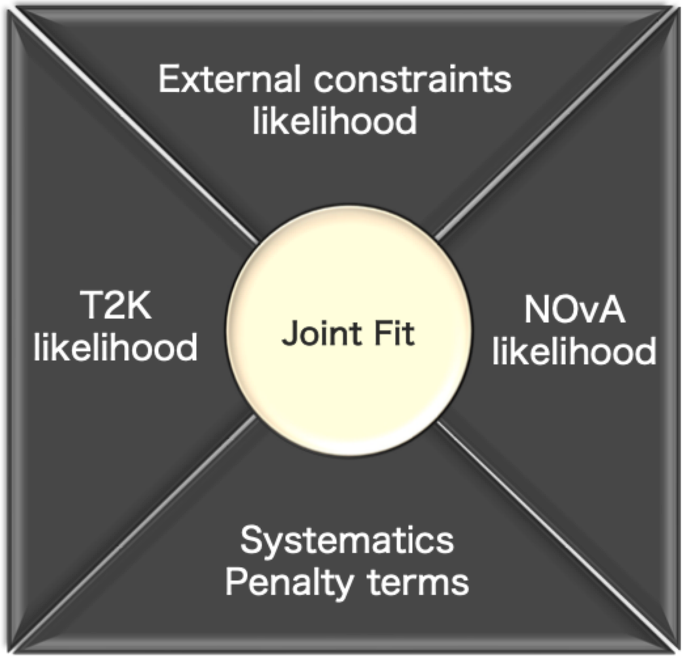
An Up-To-The-Minute Neutrino Combination: NOvA and T2K

https://indico.fnal.gov/event/62062/

Constructing the joint-analysis

3

- The joint-fit is constructed using:
  - Poisson likelihood from each experiment
  - Penalty terms from the systematics pull
  - External constraints on θ_{13}, θ_{12}, Δm^2_{21} from solar and reactor neutrino experiments
- The other experiment's likelihoods are integrated via a containerized environment.
  - Both experiments can run each other's analysis through these containers.
  - Full access to Monte-Carlo and data.



T2K   Joint Analysis Results   Zoya Vallari, Caltech   Feb 16, 2024