

Analiz ve İstatistiksel Metodlara Genel Bakış

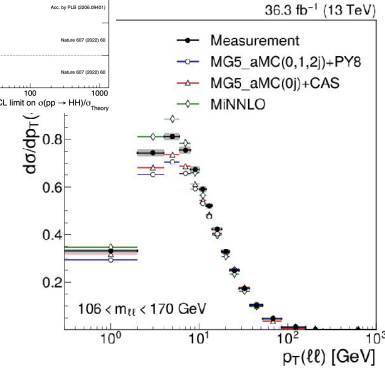
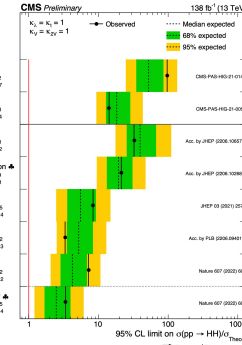
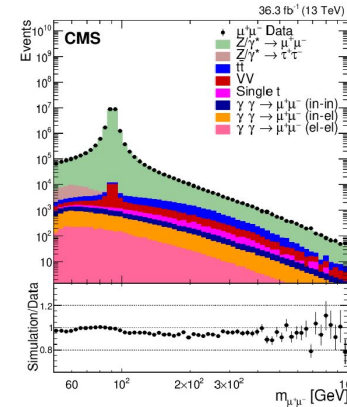
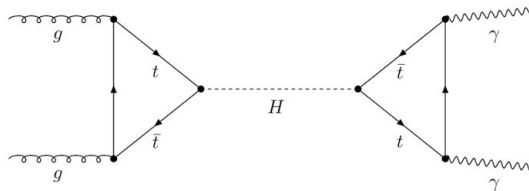
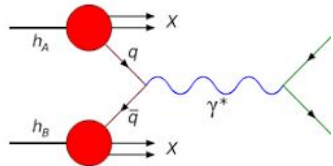
Bugra Bilin (ULB/IIHE)

Datadan sonuca nasıl ulaşılır?

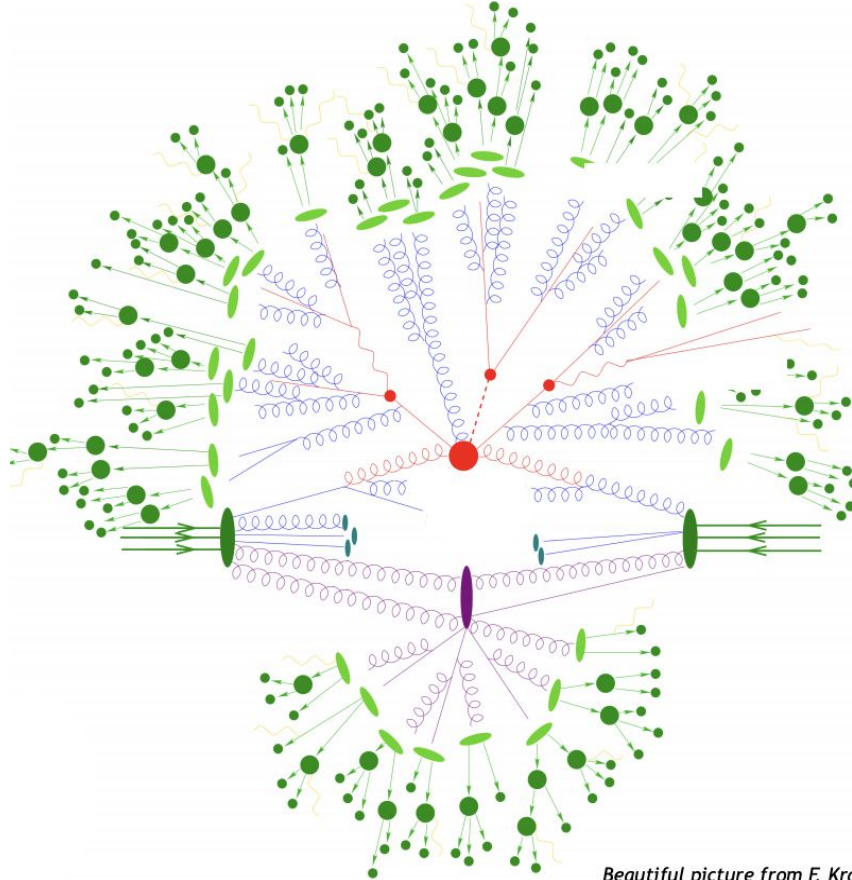
Kullanılan metodlar.

Fizik interpretasyonları

Makina Öğrenmesi ile Neler Geliştirilebilir



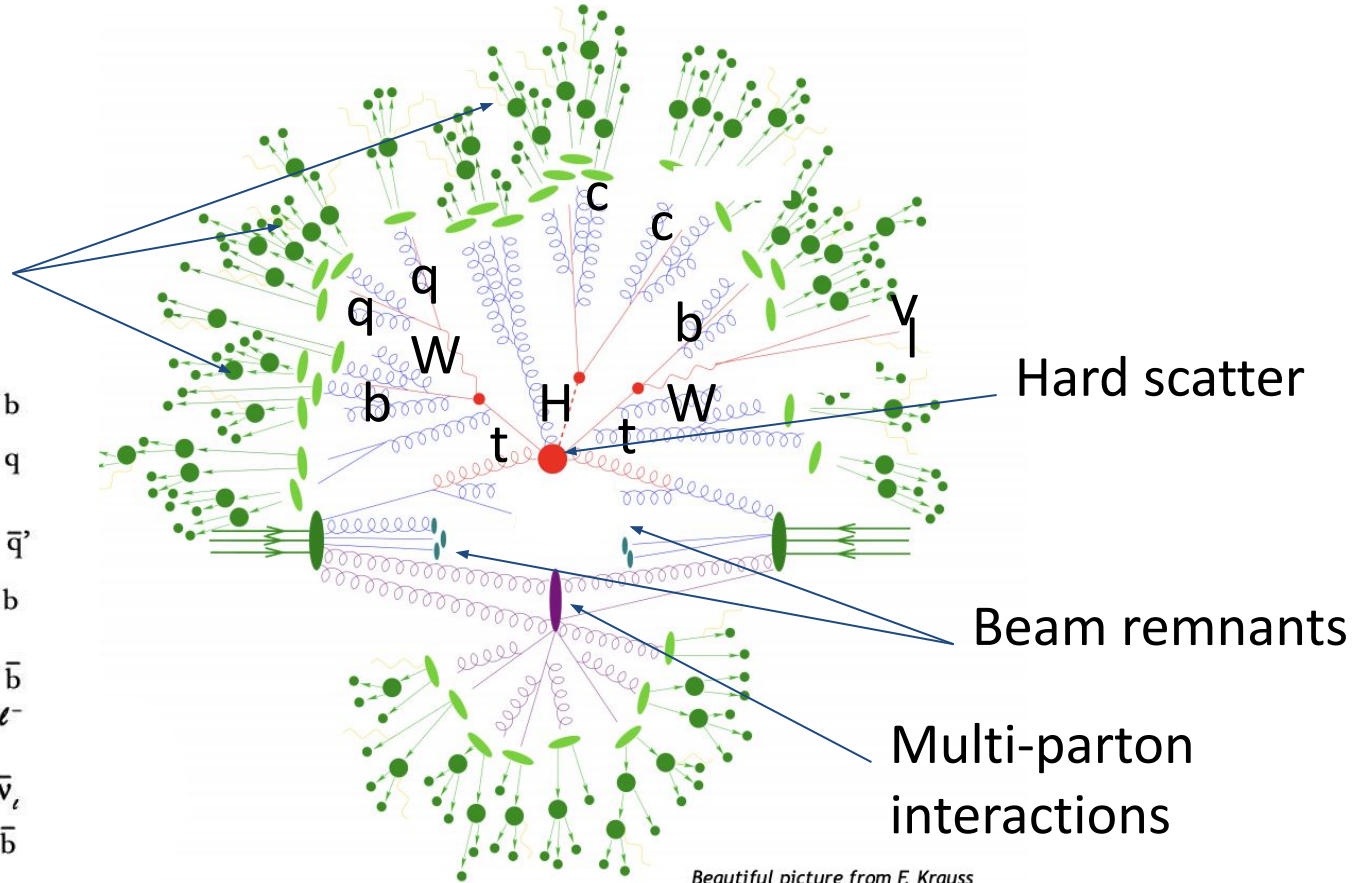
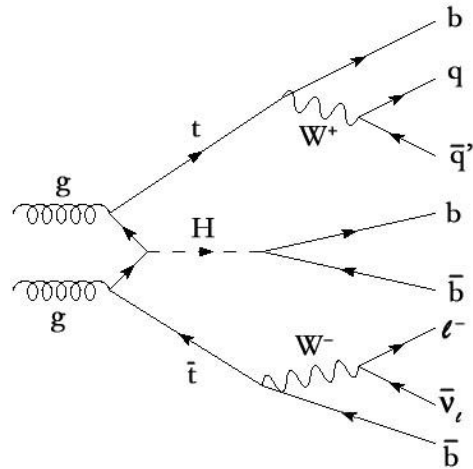
Hadron çarpıştırıcısında ne oluyor?



Beautiful picture from F. Krauss

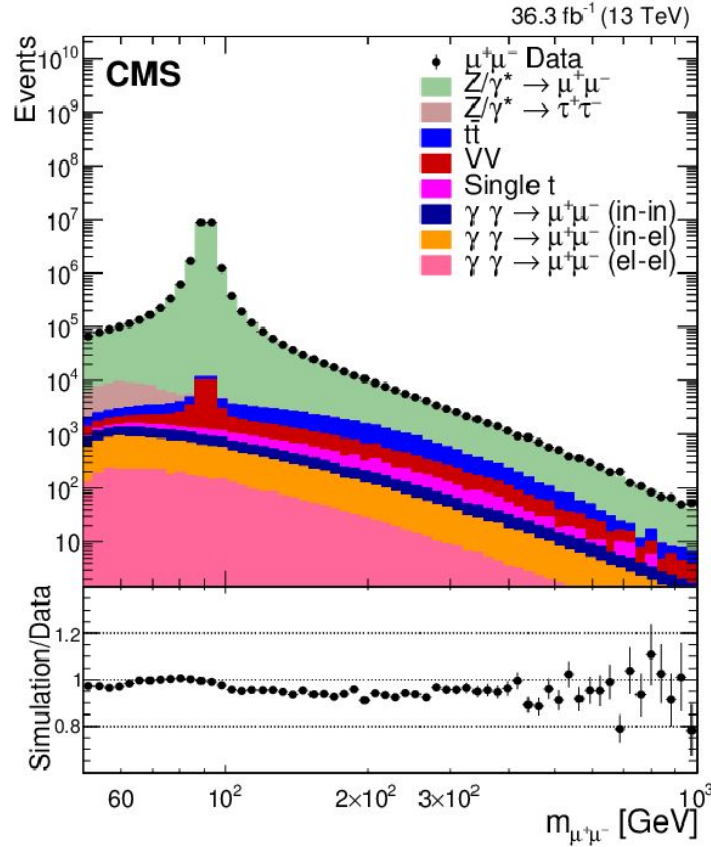
Hadron çarpıştırıcısında ne oluyor?

Parton-Shower & Hadronization



Beautiful picture from F. Krauss

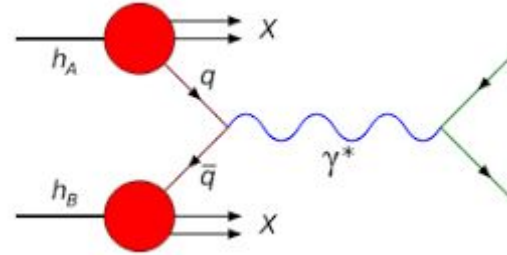
Ne ölçüyoruz: Detektör bize ne veriyor?



Data: Tüm olasılıkların toplamı

Signal: Ölçmek istediğim süreç

Ardalan (Background): Diğer süreçler.



Ardalanı temizlemem gerekiyor. Başka sorunlar?

Hatırlatma: Breit Wigner

$$\Psi(t) = \Psi(0) e^{-iEt/\hbar}$$

Lifetime

$$\Psi(t) = \frac{1}{\sqrt{\tau}} \Psi(0) e^{-iEt/\hbar} e^{-t/\tau}$$

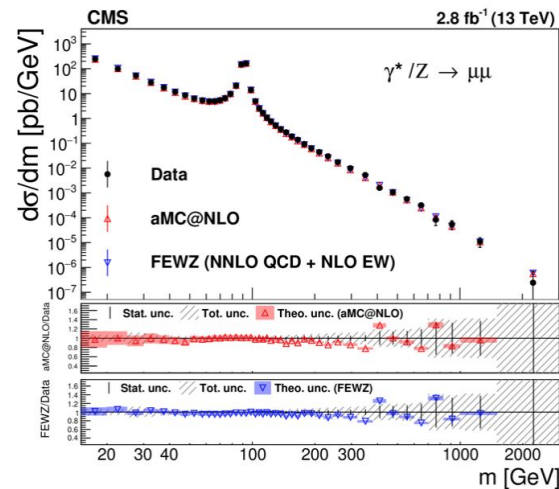
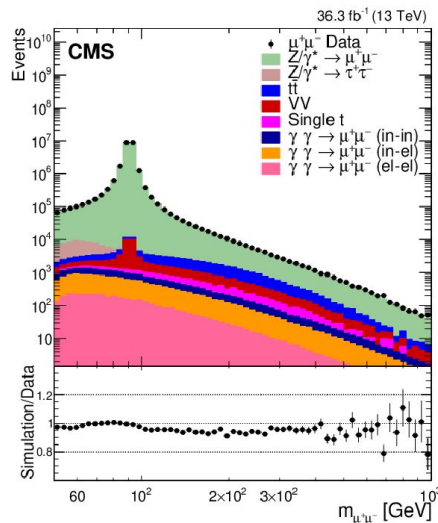
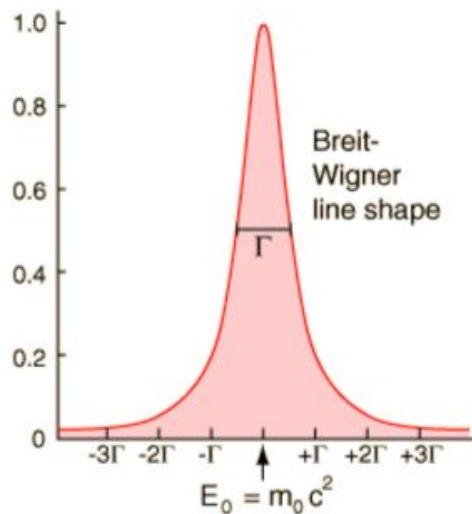
$$|\Psi(t)|^2 = \frac{1}{\tau} e^{iEt/\hbar} e^{-t/2\tau} e^{-iEt/\hbar} e^{-t/2\tau} |\Psi(0)|^2 = \frac{1}{\tau} e^{-t/\tau}$$

$$\int_0^\infty t \frac{1}{\tau} e^{-t/\tau} dt = \tau$$

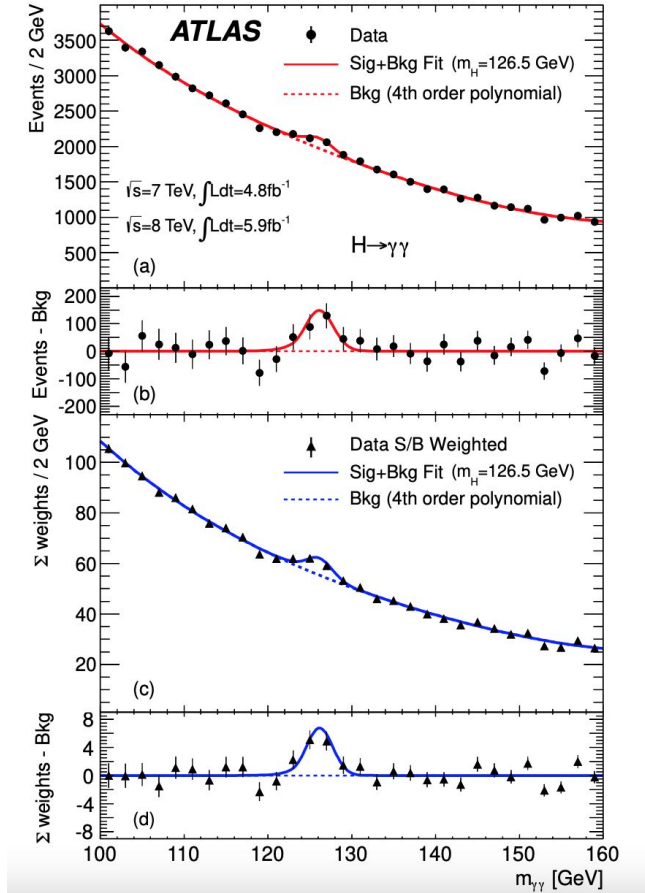
Hatırlatma: Breit Wigner

$$\begin{aligned}\chi(m) &= \int \Psi(t) e^{imt/\hbar} dt \\ &= \Psi(0) \int \sqrt{\frac{\Gamma}{\hbar}} e^{\frac{-t}{\hbar}[(i(m_0-m)-\Gamma/2)]} dt \\ &\sim \frac{\Gamma}{(m_0 - m) - i\Gamma/2}\end{aligned}$$

Enerji uzayına
Fourier transform



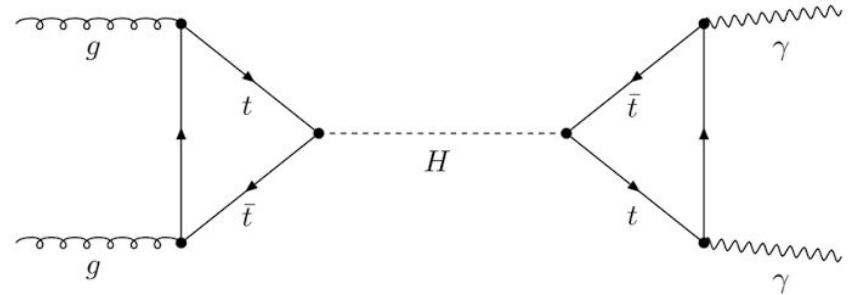
Ne ölçüyoruz: Detektör bize ne veriyor?



Higgs arıyoruz. Ardalan büyük, parametrik fit ile belirleniyor.

Excess -> Ne kadar significant? Nasıl buluş olduğuna emin oluyoruz?

Yeni fizik?



Bu derste

Detektör bize çarpışmadan gelen parçacıkları verdi.

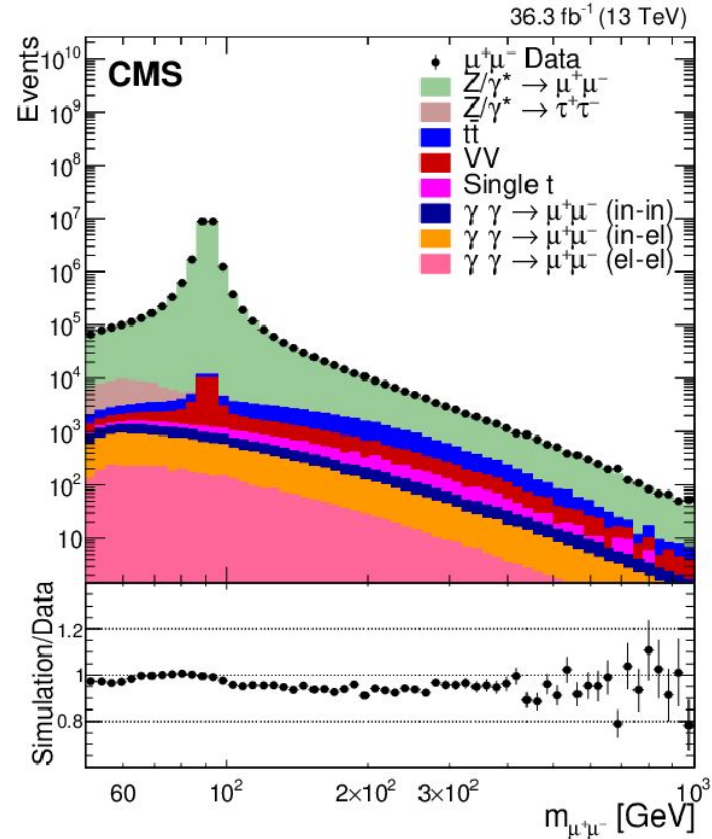
İçerisinde detektör koşullarından gelen sorunlar/ çözünürlük etkileri
Bunları nasıl düzeltiyoruz?

Ölçüm → Nasıl tesir kesiti ya da diğer fiziksel özelliklere ulaşıyoruz?

Yeni fizik var mı? Yoksa nereye kadar yok? Nasıl limit koyuyoruz?

pp → detektör → DAQ → reconstruction → Analiz → Fizik Sonucu

Drell Yan'a geri dönüş



Detektör seviyesinde dN/dm(II)
(başka kinematik dağılımlar.)

Buradan ne elde etmek istiyoruz?

$$N = \mathcal{L} \sigma$$

$$\sigma = \frac{N}{\mathcal{L}}$$

$$N_{\text{obs}} = N_{\text{sig}} + N_{\text{bkg}}$$

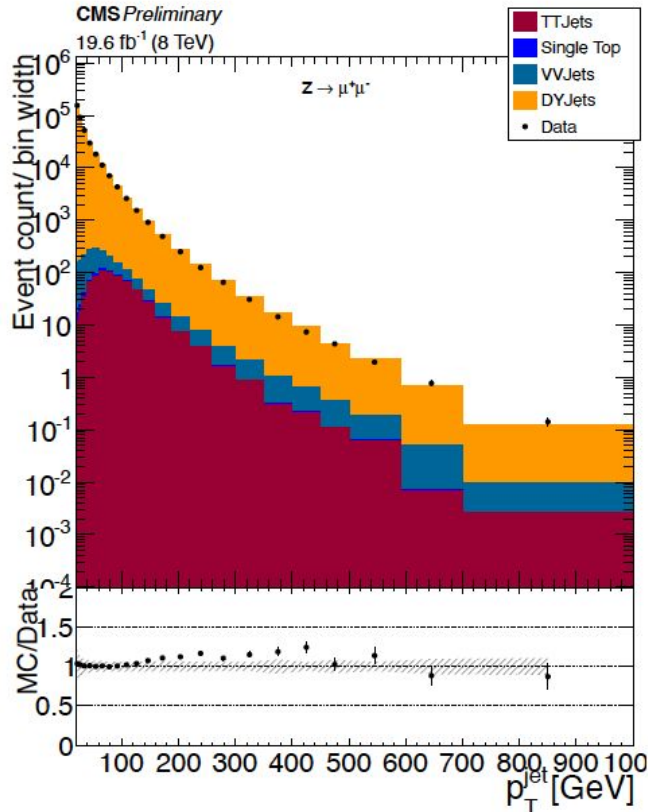
$$N_{\text{sig}} = \mathcal{L} \sigma A \epsilon$$

$$N_{\text{obs}} = N_{\text{sig}} + N_{\text{bkg}}$$

$$N_{\text{sig}} = \mathcal{L} \sigma A \epsilon$$

$$\sigma = \frac{N_{\text{obs}} - N_{\text{bkg}}}{\mathcal{L} A \epsilon}$$

Differansiyel tesir kesiti



Parton level



Particle / fiducial level

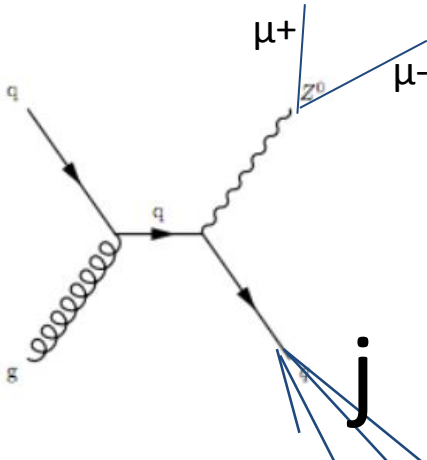
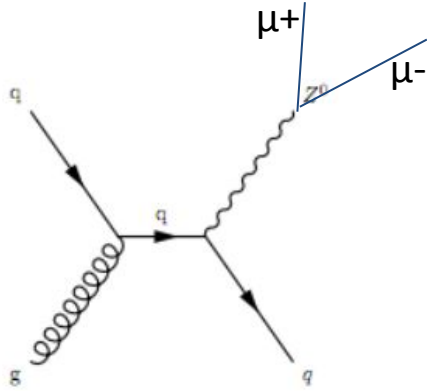


Detector / reconstruction level

$$X^{\text{reco}} \neq X^{\text{particle}}$$

Hangi seviyede bir ölçüm istiyorum?
Nereye kadar düzeltme uygulayacağım

Differansiyel tesir kesiti



Parton level

teori

Particle / fiducial level



Detector / reconstruction level

$$X^{\text{reco}} \neq X^{\text{particle}}$$

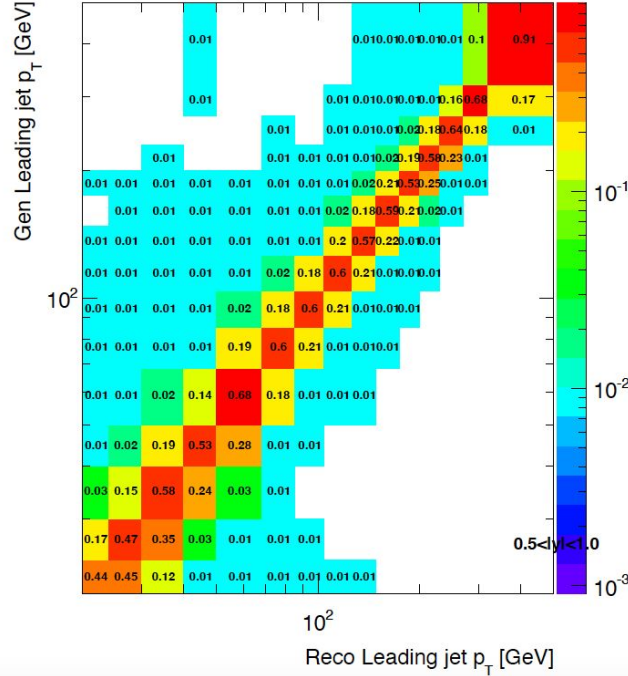
Hangi seviyede bir ölçüm istiyorum?
Nereye kadar düzeltme uygulayacağım

Differansiyel tesir kesiti

$$\frac{d\sigma_i}{dX} \sim \frac{N_i - B_i}{\mathcal{L}\epsilon_i\Delta X_i}$$

$$j_{\text{particle}} \longrightarrow i_{\text{reco}}$$

BIN MIGRATION



Parton level



Particle / fiducial level

Acceptance ve verimlilik,
migrasyonlar

Detector / reconstruction level

$$X^{\text{reco}} \neq X^{\text{particle}}$$

Hangi seviyede bir ölçüm istiyorum?
Nereye kadar düzeltme uygulayacağım

Particle Level Ölçümü

$$m_i = \sum_{j=1}^N R_{i,j} t_j$$

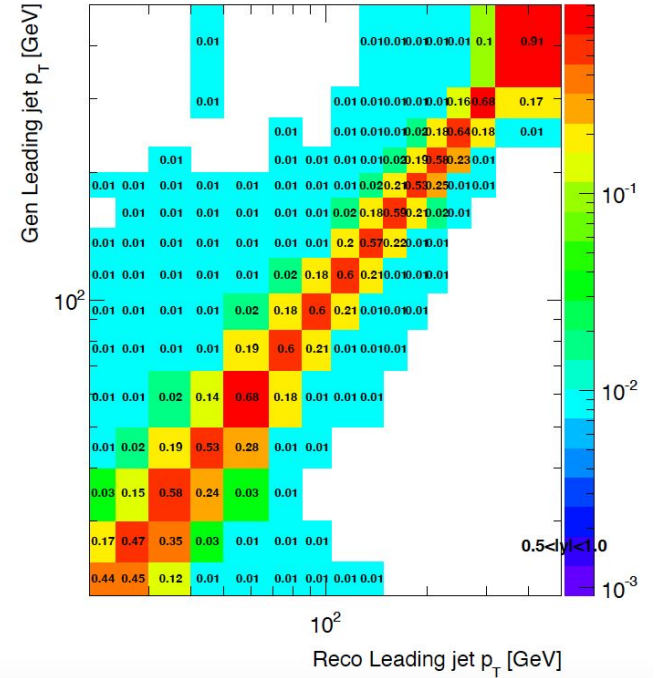
Response matrisini MC kullanarak elde edebilirim
Elimde truth (generate ettiğim) ve detektör
simülasyonum var (detaylar yarın 2. MC dersinde)

Gerisi R'nin tersini almak (!)

$t_j = R^{-1}_{j,i} m_i$ burada t kesin değil, yaklaşık..

Not: m'den b(ardalan)'ı çıkarmam lazım.

Bu analizde kolay, ardalan küçük, MC
kullanabilirim. Başka analizlerde farklı metodlar



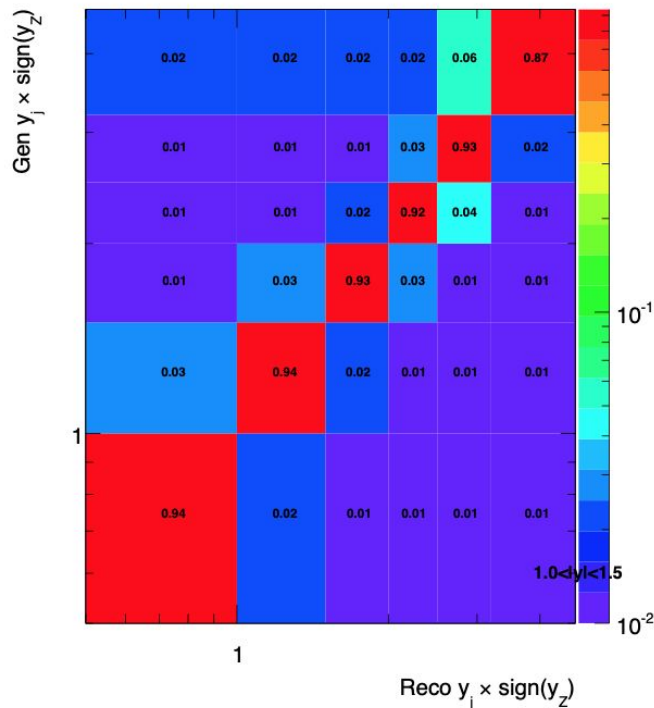
$$R_{ij} \sim P(i_{\text{reco}} | j_{\text{particle}})$$

Matrix inversion, problemler...

$$\hat{\mathbf{t}} = \mathbf{R}^{-1}(\mathbf{d} - \mathbf{b})$$

$$\delta d \longrightarrow \mathbf{R}^{-1} \delta d$$

Küçük salınımlar büyük
istatistiki hataya yol açıyor



Eğer böyle bir R varsa
matrix inversion
yeterli

Ne kadar fazla
migration varsa o
zaman matrisin
tersini almak
zorlaşıyor.

Açısal dağılımlarda resolution
daha iyi → Neden?

Regularisation

Matrisi 3 matris çarpımı olarak tanımla: $R = U \Sigma V^T$

V: Truth (gen) varyasyonları

U: Bunların detektörde nasıl görüldüğü

Σ : Detektörün bunlara sensitivitesi

$$Rv_k = \sigma_k u_k$$

σ_k elemanları ne kadar küçükse ölçüm o kadar kötü

SVD Regularisation: $\sigma_k \ll 1$ olan elemanları inversiona katma

Regularisation: Bias vs Smooth results

Güzel: Regularizasyon salınımları azaltıyor

Ama: Çok fazla regularisation $\rightarrow$ Bias, sadece belli bir kısmını kullanarak tüm ölçüm. Bunun için ortada bir değer. Testlerle sonucu doğrulamak önemli

SVD tek yöntem değil. Aşağıdaki metodları araştırın:

- $\rightarrow$ Iterative Bayesian (D'Aghostini metodu) Inversion yok
- $\rightarrow$ Tikhonov inversion

Sonuçların doğrulanması

Closure:

MCyi MC'yı ikiye böldüm
Yarısından Response'u elde ettim.
Daha küçük bir sürü 'pseudodata'
da yaratabilirim.

R yi kullanarak unfold ettim.

$$t_{MC} \xrightarrow{R} d_{MC} \xrightarrow{U} \hat{t}_{MC}$$

$$\boxed{\hat{t}_{MC} \stackrel{?}{\simeq} t_{MC}}$$

$$t_{MC} \rightarrow t_{\text{reweighted}}$$

Ya da: MC'yu dataya reweight
edip mükemmel MC elde et.

Bunu kullanarak alternate
unfolding. İkisinin farkı: unfolding
hata payı

Alternatif modeller kullanılarak
unfolding. **Soru:** Sadece unfolding
hatası mı yoksa başka etkileri de
içerebilir mi? Artık çok fazla bu
metod kullanılmıyor...

Daha detaylı bir test: Pull

Closure:

$$\hat{t}_i \stackrel{?}{\simeq} t_i^{\text{true}}$$

MCyi MC'yı ikiye böldüm

Yarısından Response'u elde ettim.

Daha küçük bir sürü 'pseudodata'
da yaratabilirim.

$$\text{pull}_i = \frac{\hat{t}_i - t_i^{\text{true}}}{\sigma(\hat{t}_i)}$$

R yi kullanarak unfold ettim.

$$d^{(1)}, d^{(2)}, \dots, d^{(N)}$$

$$d^{(k)} \xrightarrow{U} \hat{t}^{(k)}$$

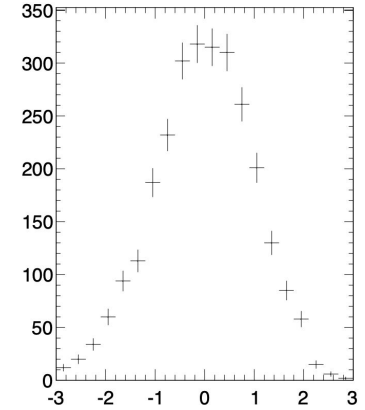


Figure B.3: Pull test for all jet y ranges combined.

Sonuç

Yorum

mean 0dan
farklı

Bias

width (>1)

Hata küçük
hesaplanmış

width (<1)

Hata büyük
hesaplanmış

Sistematik Hatalar:

İstatistiki hata payları:

$$N_{\text{data}}, \quad N_{\text{MC}}$$

$$\delta_{\text{stat}} \sim \frac{1}{\sqrt{N}}$$

Deneysel sistematik hata payları:

Lepton enerji ve momentum
Lepton reconstruction ve seçim
verimlilik hataları (Soru: Tag and Probe
method nedir?)
Luminosity, pileup corrections...
Unfolding

Ayrıca teorik hata payları hesaba
katılmakta

Son sonuca nasıl yansıyor?

Sistematik Hatalar: Nasıl belirleniyor?

Örneğin muon momentum scale:

$$p_T \rightarrow p_T(1+\delta)$$

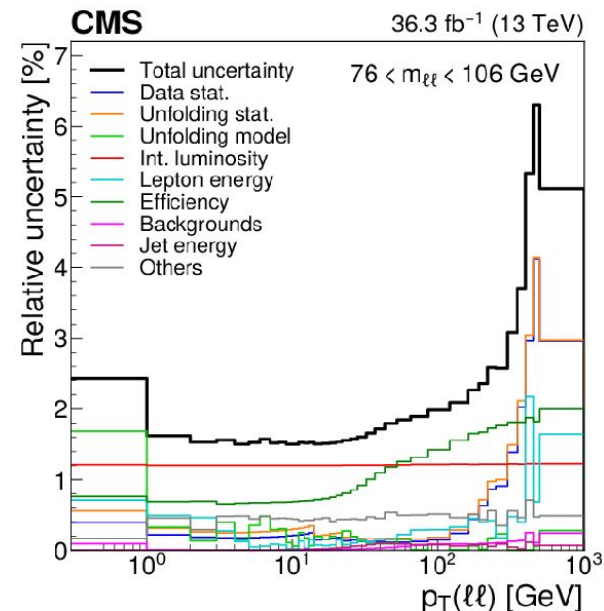
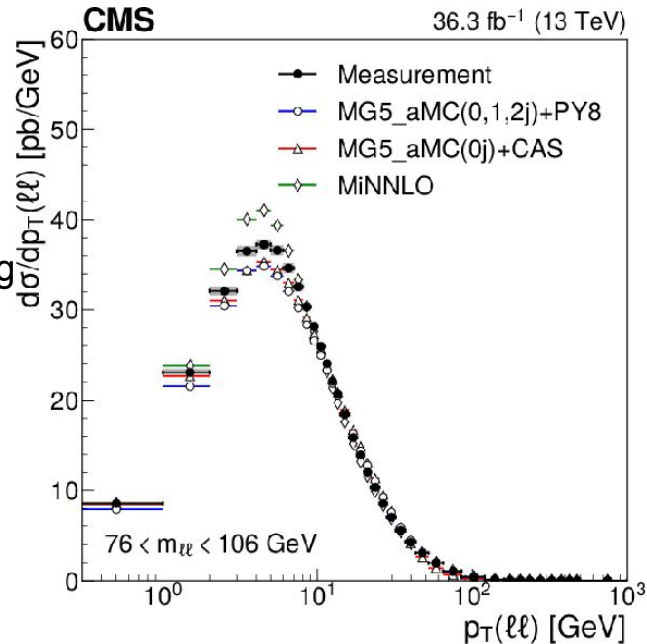
Analizi tekrar yap:

Objects $\rightarrow$ Selection $\rightarrow$ R $\rightarrow$ Unfolding

$$\hat{t}_i^0$$

$$\hat{t}_i^{\text{up}}, \quad \hat{t}_i^{\text{down}}$$

$$\Delta_i^{(s)} = \hat{t}_i^{(s)} - \hat{t}_i^0$$



Sistematik Hatalar: Covariance

Ölçüm değerleri $\rightarrow$

$$\hat{\mathbf{t}} = (\hat{t}_1, \hat{t}_2, \dots, \hat{t}_N)$$

Covariance $\rightarrow$

$$V_{ij} = \text{Cov}(t_i, t_j)$$

Bunun diagonal elementleri
uncertainty² $\rightarrow$

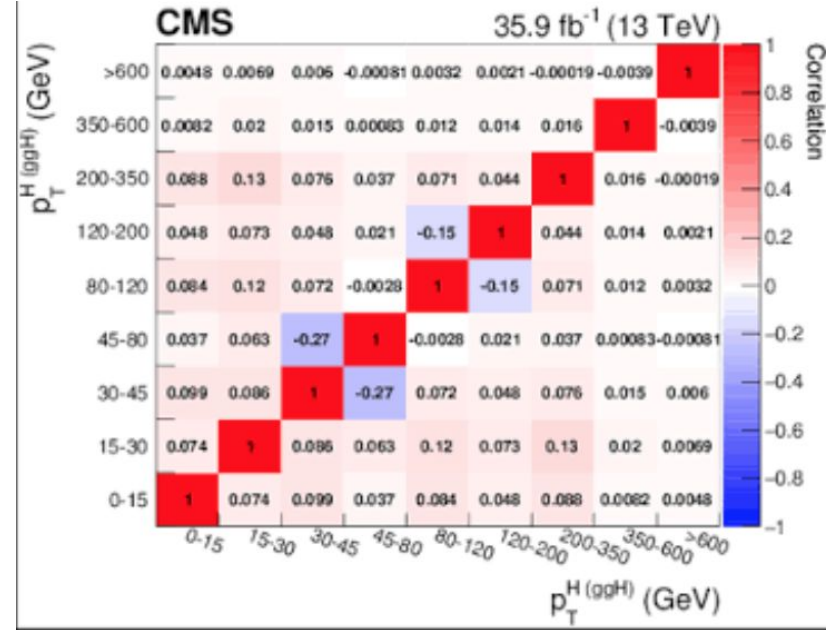
$$V_{ii} = \sigma_i^2$$

Off-diagonal elementler
0-dan farklı $\rightarrow$

$$V_{ij} \neq 0$$

Correlasyon matrisi $\rightarrow$

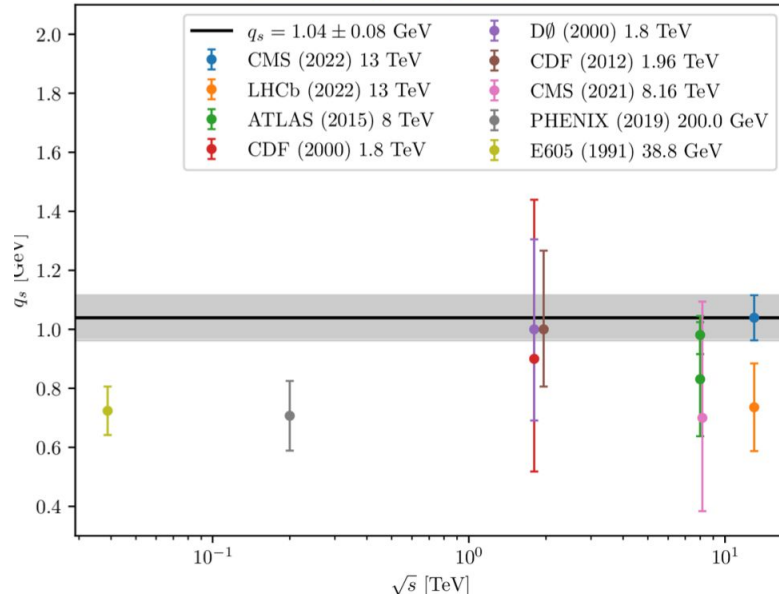
$$\rho_{ij} = \frac{V_{ij}}{\sqrt{V_{ii}V_{jj}}}$$



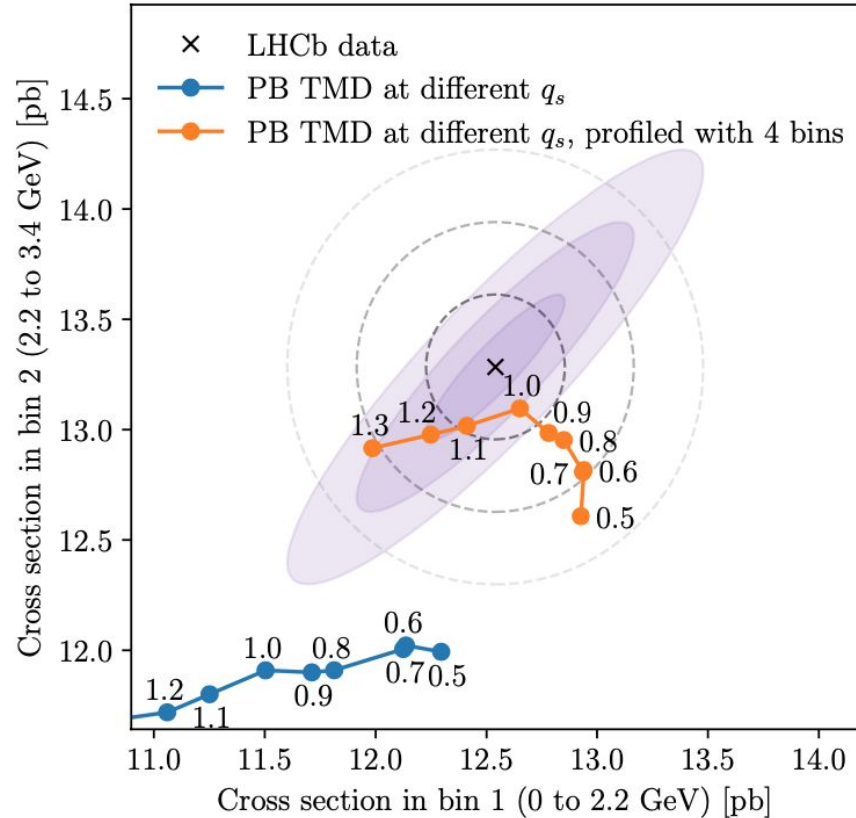
Ölçümün teori ile
karşılaştırılabilmesi için
covariance matrix önemli.

Örnek: Korelasyonları hesaba almak ya da amamak

EPJ C 84 (2024) 154 (I. Bujanja, ..., LM, ...)



Correlated or not...



Ölçümü Teoriyle Karşılaştırmak

Data (unfolded, teori level) ile teori arasındaki fark

$$\Delta = \mathbf{d} - \mathbf{t}$$

Eğer correlation olmasaydı

$$\chi^2 = \sum_i \frac{(d_i - t_i)^2}{\sigma_i^2}$$

Ama tüm binler bağlı (correlated)

$$\chi^2 = (\mathbf{d} - \mathbf{t})^T V^{-1} (\mathbf{d} - \mathbf{t})$$

ÖNEMLİ: Bir teorist bizim unfolded ölçümümüzü yeni bir model ile karşılaştırmak istiyorsa sadece noktalar (d) ve error bar'lar yeterli değil.

Hepdata'ya bakınız. Ölçümlerle ilgili ne verilmiş?



Repository for publication-related High-Energy Physics data

Parantez: Ardalan Belirleme

Biraz önceki ölçüm örneğinde ardaan küçüktü. Peki ardaan büyükse?

$$N_{\text{sig}} = N_{\text{obs}} - N_{\text{bkg}}$$

N_{bkg} 'yi nereden biliyorum?

Az önceki örnekte MC'dan aldık:

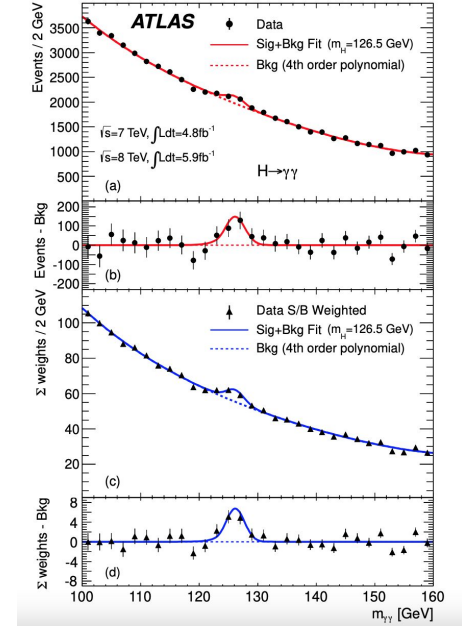
$$N_{\text{bkg}}^{MC}$$

ilk alternatif: Control Bölgesinden almak. Örneğin higgs

$$\text{Data}_{CR} \rightarrow N_{\text{bkg}}^{SR}$$

Tek bir variable bana bu ayrımı yapamayabilir. O durumda:

ABCD



ABCD

İki variable buldum, x ve y

	y_{low}	y_{high}
x_{low}	D	C
x_{high}	B	A

A bölgesi sinyalimin olduğu bölge

$$\frac{N_A}{N_B} = \frac{N_C}{N_D}$$

A bölgesindeki ardaan beklentisi

$$N_A^{\text{pred}} = \frac{N_B N_C}{N_D}$$

PROBLEM: Korelasyonlar

İki parametre birbirinden tamamen bağımsız değil. Soldaki $k \neq 1$ olmayabilir.

ABCD'nin bütün problemi iki iyi variable bulmak.

İki variable: signal-background'u iyi ayıracak, background için birbirinden bağımsız olacak.

Bu ikisini aynı anda bulmak zor.

Yapay zeka yardım edebilir mi?

$$\kappa = \frac{N_A N_D}{N_B N_C}$$

$$\kappa = 1$$

ABCD $\rightarrow$ ABCDisCo(TEC)

isim bulmakta da yaratıcı olabiliyoruz :)

ABCD'de x ve y'yi ben buldum

Tüm variablelar kullanılarak

$$\text{event variables} \xrightarrow{\text{NN}} (f_1, f_2)$$

İki bağımsız variable bulabilir.

Bu bana istediğim sinyal/ardalan ayırmasını verebilir.

ABCDisCo metodu bunu yapıyor

ABCDisCo decorrelation'ı optimize ediyor.

ABCDisCoTEC analizin asıl ihtiyacı olan ABCD closure'ı loss'a doğrudan sokuyor.

arXiv

High Energy Physics – Experiment

[Submitted on 10 Jun 2025]

Machine learning method for enforcing variable independence in background estimation with LHC data: ABCDisCoTEC

CMS Collaboration

A novel solution is presented for the problem of estimating the backgrounds of a signal search using observed data while simultaneously maximizing the sensitivity of the search to the signal. The “ABCD method” provides a reliable framework for background estimation by partitioning events into one signal-enhanced region (A) and three background-enhanced control regions (B, C, and D) via two statistically independent variables. In practice, even slight correlations between the two variables can significantly undermine the method’s performance. Thus, choosing appropriate variables by hand can present a formidable challenge, especially when background and signal differ only subtly. To address this issue, the ABCD with distance correlation (ABCDisCo) method was developed to construct two artificial variables from the output scores of a neural network trained to maximize signal–background discrimination while minimizing correlations using the distance correlation measure. However, relying solely on minimizing the distance correlation can yield undesirable characteristics in the resulting distributions, which may compromise the validity of the background prediction obtained using this method. The ABCDisCo training enhanced with closure (ABCDisCoTEC) method is introduced to solve this issue by directly minimizing the nonclosure, expressed as a dedicated differentiable loss term. This extended method is applied to a data set of proton–proton collisions at a center-of-mass energy of 13 TeV recorded by the CMS detector at the CERN LHC. Additionally, given the complexity of the minimization problem with constraints on multiple loss terms, the modified differential method of multipliers is applied and shown to greatly improve the stability and robustness of the ABCDisCoTEC method, compared to grid search hyperparameter optimization procedures.

Zyi ölçtük, peki W?

$$Z \rightarrow \mu^+ \mu^-$$

İki müonu da görebiliyorum/ölçebiliyorum

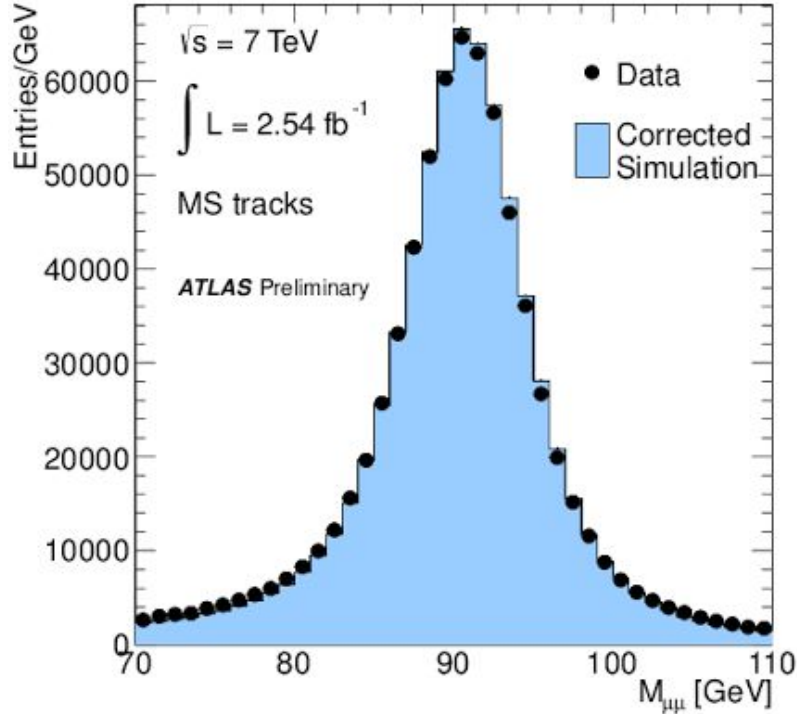
$$m_{\mu\mu}^2 = (p_{\mu^+} + p_{\mu^-})^2$$

Neutrino ları göremiyorum $\rightarrow$ MET

$$W \rightarrow \mu \nu$$

Z ölçümünde doğrudan detektör seviyesinden unfolding yapabilmıştık. W kütlesi için farklı bir strateji gerekli

mW



Z kütlesini çok iyi görebiliyorum

Bunu kullanarak W ölçümü için muon enerji ve resolution kalibrasyonu yapabilirim.

Ayrıca Z olaylarını W metodunu test etmek için kullanabilirim.

mW

$$W \rightarrow \mu\nu$$

mW direk göremiyorum ama

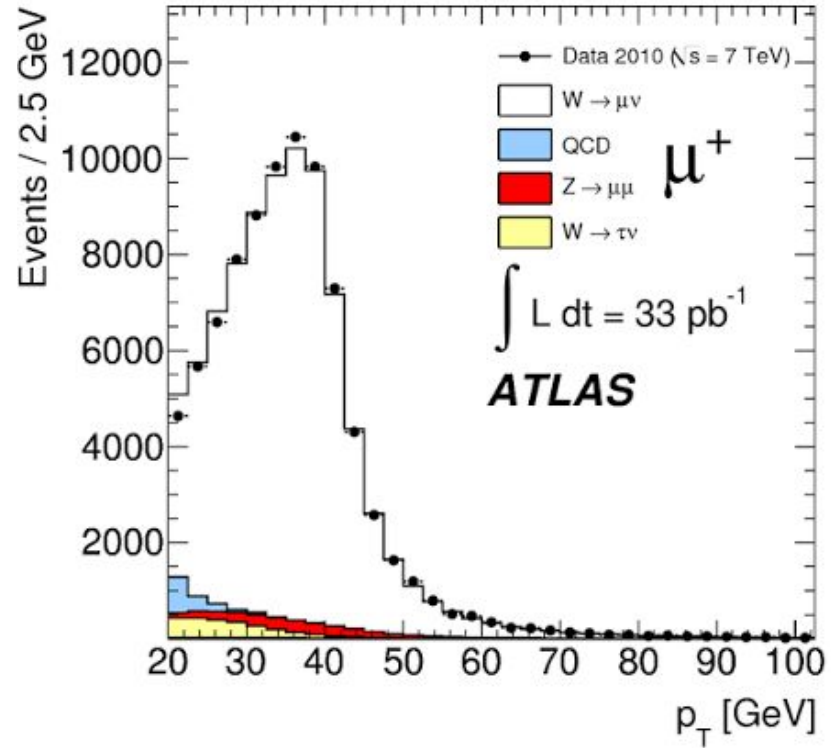
p_T^μ ölçebiliyorum

Detektör frame'inde

$$p^\mu \simeq \frac{m_W}{2}$$

Farklı kütle değerleri p_T 'yi etkiler

$m_W \rightarrow p_T^\mu$ shape



mW: Unfold yerine forward fit

Burada detector effect'lerini tersine çevirmiyorum.

Farklı m_W hipotezlerini teoriden ileri doğru geçiriyorum ve hangisi data'yı en iyi açıklıyor diye soruyorum

Farklı m_W değerleri için hipotezler üretip dağılım elde et: $\nu_i(m_W)$

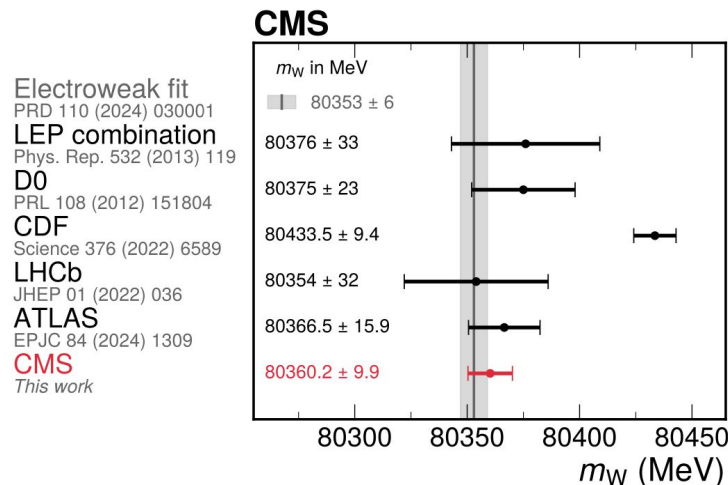
Datada gözlemlediğim n_i

Zor bir fit: Detaylar için CMS
Combine tool'a bakabilirsiniz.

$$L(m_W, \theta) = \prod_i \text{Pois}(n_i | \nu_i(m_W, \theta)) \prod_k \pi_k(\theta_k)$$

θ : systematic uncertainties

$$m_W = 80360.2 \pm 9.9 \text{ MeV}$$



Likelihood: Biraz detaylı bakalım

Observed:

$$n$$

Expected:

$$\nu(\mu, \theta)$$

Poisson:

$$P(n|\nu) = \frac{\nu^n e^{-\nu}}{n!}$$

Likelihood:

$$L(\mu, \theta) = P(\text{data}|\mu, \theta)$$

Burada:

μ = Parameter of Interest

θ = Nuisance Parameters

Physics parameter:

$$\mu$$

Örneğin:

$$m_W, \quad \mu_H, \quad \sigma_{\text{BSM}}$$

Ama prediction ayrıca:

$$\theta = (\theta_{\text{lumi}}, \theta_{\text{scale}}, \theta_{\text{PDF}}, \theta_{\text{bkg}}, \dots)$$

parametrelerine bağlı.

$$\nu_i = \nu_i(\mu, \theta)$$

Sistematikler nuisance parametresi

Higgs Search, nasıl?

Data:

$$H \rightarrow \gamma\gamma$$

Signal:

narrow resonance around

$$m_H \simeq 125 \text{ GeV}$$

Background:

smoothly falling spectrum.

$$\boxed{\text{Signal} + \text{Background}}$$

$$\mu_H = \frac{\sigma_H^{\text{measured}}}{\sigma_H^{\text{SM}}}$$

$$\mu_H = 0$$

Higgs yok

$$\mu_H = 1$$

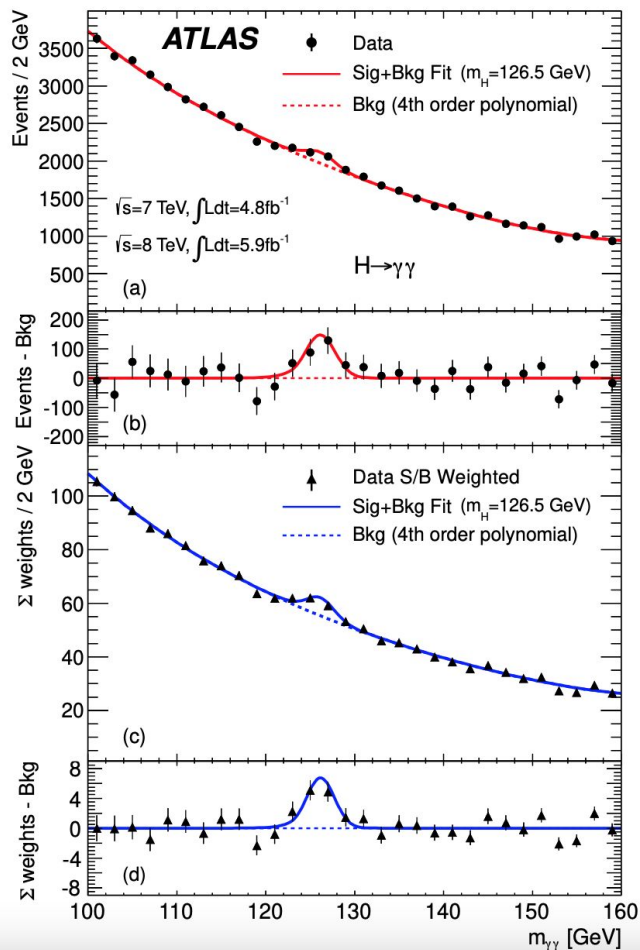
Higgs SM'in
beklediği gibi

$$L(\mu_H, \theta)$$

Burada W mass yerine fit bana

bu oranı verecek $\hat{\mu}_H \pm \Delta\mu_H$

Higgs Search, how?



$$H_0 : \mu = 0$$

$$q_0 = -2 \ln \lambda(0)$$

$$p_0 = P(q_0 \geq q_0^{obs} | H_0)$$

$$p_0 \rightarrow Z$$

$$Z = 5\sigma$$

BG only hypothesis'in excessi açıklama
olasılığı $>5\sigma \rightarrow$ Discovery!!!

BSM: Excess yok, ne öğrendim?

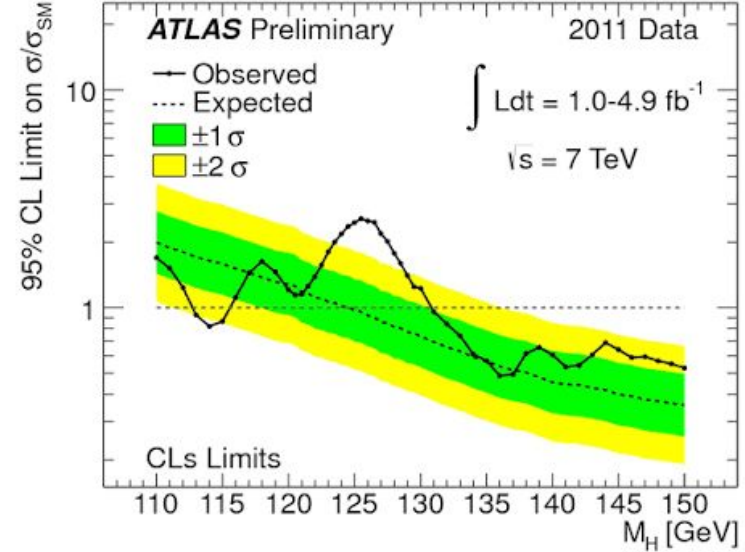
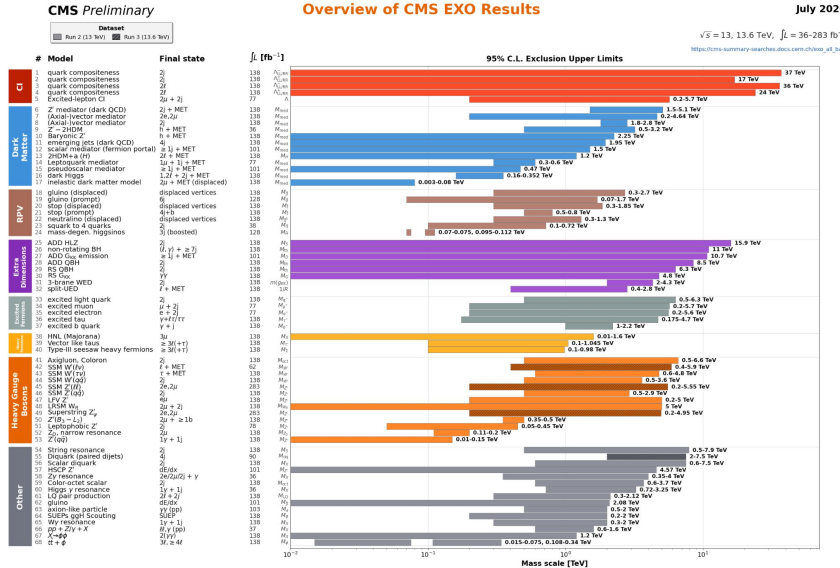
Örneğin:

$$pp \rightarrow X \rightarrow \ell^+ \ell^-$$

Hangi modeller iki lepton verebilir?

Data'da significant excess yok.

Buluş yok, o da bir fizik sonucu
Teori nereye kadar datayla uyumlu
değil?



Makina Öğrenmesi ile Ne geliştirebilirim?

Bugünkü zincire geri dönelim:

Data $\rightarrow$ Objects $\rightarrow$ Selection $\rightarrow$ Observable $\rightarrow$ Inference

ML kullanılabilir:

Reconstruction
Signal/background separation
Background estimation
Unfolding
Sensitivity optimization

Cut based yerine optimised selection

Cut-based:

$$p_T > 30, \quad m > 100, \quad \Delta R < 1.5$$

ML:

$$x = (p_T, m, \eta, \Delta R, \dots)$$

$\Downarrow$

$$\boxed{f(x) = y_{\text{ML}}}$$

Sonra:

$$y_{\text{ML}} \rightarrow L(\mu, \theta)$$

Training:

$$p_{\text{MC}}(x)$$

Data:

$$p_{\text{data}}(x)$$

Eğer:

$$p_{\text{MC}}(x) \neq p_{\text{data}}(x)$$

ML simulation'ın yanlış modellediği özelliği öğrenebilir.

Bir classifier S/B'yi yüzde 20 iyileştirip systematic uncertainty'yi iki katına çıkarıyorsa bize yardımcı olmamış olabilir.

Datadan fizik sonucuna

pp → detektör → DAQ → reconstruction → **Analiz** → **Fizik Sonucu**

Selection
Background
Detector corrections
Systematics
Statistical inference
ML

$$\left[\frac{d\sigma}{dX}, \quad m_W, \quad \mu_H, \quad Z, \quad \sigma_{95} \right]$$

1. Detector-level distribution $\neq$ physics result
2. $d = Rt + b \Rightarrow$ Unfolding is an inverse problem
3. Measurement = value + uncertainty + correlations
4. Systematics belong inside the analysis
5. $L(\mu, \theta)$ connects data to physics parameters
6. ML should improve physics sensitivity