Universidade do Estado do Rio de Janeiro
Centro de Tecnologia e Ciências
Instituto de Física Armando Dias Tavares

Gabriel Moreira da Silva Campos

Busca por matéria escura no CMS/LHC: um estudo de implementação de métodos de aprendizado de máquina e aplicação de fatores de correção para jatos de quark bottom

Rio de Janeiro
2023

Gabriel Moreira da Silva Campos

Busca por matéria escura no CMS/LHC: um estudo de implementação de métodos de aprendizado de máquina e aplicação de fatores de correção para jatos de quark bottom

Dissertação apresentada, como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-Graduação em Física, da Universidade do Estado do Rio de Janeiro.

Orientadora: Prof. Dra. Helena Brandão Malbouisson
Coorientador: Prof. Dr. Dilson de Jesus Damião

Rio de Janeiro
2023

CATALOGAÇÃO NA FONTE
UERJ/ REDE SIRIUS / BIBLIOTECA CTC/D

C198b Campos, Gabriel Moreira da Silva.
Busca por matéria escura no CMS/LHC: um estudo de implementação de métodos de aprendizado de máquina e aplicação de fatores de correção para jatos de quark bottom / Gabriel Moreira da Silva Campos. – 2023.
120 f. : il.

Orientadora: Helena Brandão Malbouisson.
Coorientador: Dilson de Jesus Damião.
Dissertação (mestrado) - Universidade do Estado do Rio de Janeiro, Instituto de Física Armando Dias Tavares.

1. Matéria escura (Astronomia) – Teses. 2. Aprendizado do computador – Teses. 3. Modelo padrão (Física nuclear) – Teses. 4. Solenóide de múon compacto – Teses. 5. Grande colisor de hádrons (França e Suíça) – Teses. I. Malbouisson, Helena Brandão (Orient.). II. Damião, Dilson de Jesus (Coorient.). III. Universidade do Estado do Rio de Janeiro. Instituto de Física Armando Dias Tavares. IV. Título.

CDU 524.5-46

Bibliotecária: Teresa da Silva CRB7/5209

_____ _____

Assinatura Data

Gabriel Moreira da Silva Campos

Busca por matéria escura no CMS/LHC: um estudo de implementação de métodos de aprendizado de máquina e aplicação de fatores de correção para jatos de quark bottom

Dissertação apresentada, como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-Graduação em Física, da Universidade do Estado do Rio de Janeiro.

Aprovada em: 28 de fevereiro de 2023.

Banca Examinadora:

_____
Prof. Dra. Helena Brandão Malbouisson (Orientadora)
Instituto de Física Armando Dias Tavares – UERJ

_____
Prof. Dr. Dilson de Jesus Damião (Coorientador)
Instituto de Física Armando Dias Tavares – UERJ

_____
Prof. Dra. Eliza Melo da Costa
Instituto de Física Armando Dias Tavares – UERJ

_____
Prof. Dr. Marcelo Santos Guimarães
Instituto de Física Armando Dias Tavares – UERJ

_____
Prof. Dr. Carsten Hensel
Centro Brasileiro de Pesquisas Físicas

_____
Prof. Dr. Gilson Correia Silva
Deutsches Elektronen-Synchrotron

Rio de Janeiro

2023

DEDICATÓRIA

Ao amor da minha vida, Ângela Azevedo.

AGRADECIMENTOS

Agradeço à minha família que independente de minhas escolhas sempre me apoiou e me motivou a seguir com todo seu amor.

À minha noiva, Ângela que acima de tudo me manteve no caminho certo para finalizar esse trabalho. Você é minha maior inspiração.

Aos meus orientadores, Helena e Dilson que como excelente profissionais me guiaram e tiverem a paciência necessária para que eu concluísse esse trabalho.

Às amizades que construí dentro da faculdade, Victor, Raphael e Matheus, pelas boas conversas e incentivos ao longo da pós-graduação.

A todos os integrantes da colaboração dessa pesquisa pelo conhecimento adquirido ao longo dessa caminhada, em especial para o Gilson que certamente atuou como um terceiro orientador ao meu trabalho.

Por fim, agradeço a todos os funcionários e professores do departamento que contribuíram, direta ou indiretamente, para minha formação.

A ciência, meu rapaz, é feita de erros, mas de erros que é bom cometer, pois levam, pouco a pouco, à verdade.

Júlio Verne

RESUMO

CAMPOS, G. M. S. *Busca por matéria escura no CMS/LHC*: um estudo de implementação de métodos de aprendizado de máquina e aplicação de fatores de correção para jatos de quark bottom. 2023. 120 f. Dissertação (Mestrado em Física) – Instituto de Física Armando Dias Tavares, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2023.

Nesse trabalho é apresentado um estudo da implementação de algoritmos de aprendizado de máquina e correção de eventos de simulação com a presença de jatos provenientes do quark bottom na busca por matéria escura fermiônica produzida através do processo de decaimento de um bóson de Higgs (H) pesado originado da quebra de simetria espontânea de dois dupletos de Higgs. O estado final do processo $\bar{b}bZ(\rightarrow \ell\bar{\ell})+\not{E}_T(a \rightarrow \chi\bar{\chi})$ é sondado utilizando todos os dados disponíveis no Run-2 do LHC/CMS. A presença de jatos provenientes de quark bottom no estado final tornam obrigatória a correção dos eventos de simulação de modo que exista uma boa concordância com os dados. Devido à baixa seção de choque do sinal, a performance e discriminantes dos algoritmos XGBoost e *Multi Layer Perceptron* foram comparados para melhor determinação da sensibilidade do sinal.

Palavras-chave: Física Experimental de Altas Energias. Experimento CMS. Matéria Escura. Aprendizado de Máquina.

ABSTRACT

CAMPOS, G. M. S. *Search for dark matter at CMS/LHC*: a study on the implementation of machine learning methods and application of correction factors for bottom quark jets. 2023. 120 f. Dissertação (Mestrado em Física) – Instituto de Física Armando Dias Tavares, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2020.

In this work, a study of the implementation of machine learning algorithms and correction of simulation events with the presence of quark bottom jets in the search for fermionic dark matter produced through the decay process of a heavy Higgs boson (H) from the spontaneous symmetry breaking of two Higgs doublets is presented. The process final state $\bar{b}bZ(\rightarrow \ell\bar{\ell}) + \not{E}_T(a \rightarrow \chi\bar{\chi})$ is probed using all available data in Run-2 from LHC/CMS. The presence of quark bottom jets in the final state makes it mandatory to correct the simulation events so that there is good agreement between data and Monte Carlo. Due to the low cross section of the signal, the performance and discriminants of the XGBoost and *Multi Layer Perceptron* algorithms were compared to better determine the signal sensitivity.

Keywords: Experimental High Energy Physics. CMS Experiment. Dark Matter. Machine Learning.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

INTRODUÇÃO

A matéria escura (*Dark Matter*, DM) é um problema não resolvido na física contemporânea, alocado na interface entre a física de partículas e a cosmologia. Nesse trabalho iremos analisar dados de eventos do *Compact Muon Solenoid* (CMS) e simulação de Monte Carlo (MC) nos quais buscamos encontrar um férmion de Dirac χ, sendo uma pequena extensão do Modelo Padrão (MP) classificado como um *Weakly Interacting Massive Particle* (WIMP). Nesse contexto é introduzido o modelo *two Higgs doublet model* (2HDM) (1) como uma realização natural de uma interface entre DM e o MP renormalizável e invariante de calibre tal que o setor do Higgs é expandido com um segundo dubleto de Higgs.

Como sugerido em (2) essa busca é inédita no *Large Hadron Collider* (LHC) e a interface pseudoescalar entre o setor visível do MP e setor escuro além do MP pode ser sondado através do processo $\bar{b}bZ(\to \bar{\ell}\ell) + \not{E}_T$, onde identificamos o estado final com dois jatos provenientes do quark bottom ($\bar{b}b$), dois léptons de cargas opostas ($\bar{\ell}\ell$) e energia faltante ($\not{E}_T$). A análise *blind*[1] é feita com o uso de aprendizado de máquina supervisionado como um caminho para o cálculo das regiões de exclusão.

A análise está sendo feita em uma colaboração entre o Centro Brasileiro de Pesquisas Físicas (CBPF), o Deutsches Elektronen-Synchrotron (DESY) e a Universidade do Estado do Rio de Janeiro (UERJ). A abordagem adotada segue os seguintes passos: a determinação dos *triggers*, seleção do sinal, determinação dos *backgrounds* e regiões de controle, aplicação de correções nos eventos de simulação, cálculo das incertezas sistemáticas, construção de discriminante entre sinal e *background* utilizando aprendizado de máquina e por fim a determinação da região de exclusão. O objetivo desse trabalho é centrado na seleção do sinal utilizando algoritmos de aprendizado máquina supervisionados e a implementação da reponderação dos eventos de simulação utilizando um algoritmo de *b-tagging* para identificação de jatos provenientes do quark bottom.

O primeiro capítulo dessa dissertação tem por objetivo apresentar o Modelo Padrão das Partículas Elementares e conceitos além, uma breve discussão de como o problema de massa faltante realiza uma interdisciplinaridade natural entre a Cosmologia e a Física de Partículas. Ainda no primeiro captítulo será feita uma breve descrição do modelo 2HDM que é peça fundamental para apresentação do processo físico, estudado na análise, capaz de produzir matéria escura fermiônica. O segundo capítulo é aborda o LHC e seus sub-detectores, em especial o experimento CMS, o aparato experimental responsável por produzir os dados utilizados nessa análise. O terceiro capítulo é responsável

[1] Estratégia de análise de dados utilizada para remover o *bias* do analista na definição de critério de seleção dos eventos sem favorecer os eventos na região em que o Sinal é esperado.

por explicar os algoritmos de aprendizado de máquina utilizados na análise. Por fim, o capítulo 4 apresenta as amostras de dados utilizados, os métodos utilizados na análise e os resultados obtidos.

1 MODELO PADRÃO DAS PARTÍCULAS ELEMENTARES E ALÉM

O Modelo Padrão da Física de Partículas é uma teoria que explica três das quatro interações fundamentais conhecidas na natureza: eletromagnetismo, interação fraca e interação forte. A teoria descreve a natureza como sendo formada por partículas fundamentais chamadas férmions que interagem através da troca de partículas portadoras de força chamadas bósons.

No entanto, ainda que o Modelo Padrão seja uma das teorias de maior sucesso na história humana, ela não é definitiva e tem suas dificuldades. A interação gravitacional descrita através da Teorai da Relatividade Geral (RG) não está integrada no MP, pois, diferente das outras interações fundamentais que são descritas com ajuda da Teoria Quântica de Campos através do transporte de informação entre partículas mediadores, a gravidade descrita na RG não é uma força e sim uma propriedade do espaço-tempo e o arcabouço matemático é incompatível com o MP. Várias tentativas foram feitas de descrever a gravidade em física de partículas como um modelo além do MP, o gráviton, uma partítula mediadora hipotética, é uma delas, porem não foi observada experimentalmente. A existência de matéria indetectável através do espectro eletromagnético nas galáxias, chamada de Matéria Escura, é outro problema que o Modelo Padrão não se propõe a explicar.

Esses e outros problemas em aberto tornam necessário (a.) propor uma nova teoria fundamental capaz de explicar problemas já resolvidos e os ainda em aberto ou (b.) considerar a teoria atual como uma teoria não fundamental de baixas energias e construir uma teoria mais completa a partir desta adicionando correções de maior ordem. Nesse trabalho buscamos estudar experimentalmente o segundo caso, na busca por matéria escura fermiônica capaz de interagir com os objetos definidos no Modelo Padrão a partir de uma extensão na teoria. Os presentes capítulos trazem um resumo dos principais pontos acerca do Modelo Padrão e da Matéria Escura e sua interdisciplinaridade entre a Cosmologia e a Física de Partículas, bem como uma breve discussão sobre o modelo *Two Higgs Doublet Model* e o processo físico capaz de produzir matéria escura fermiônica (χ).

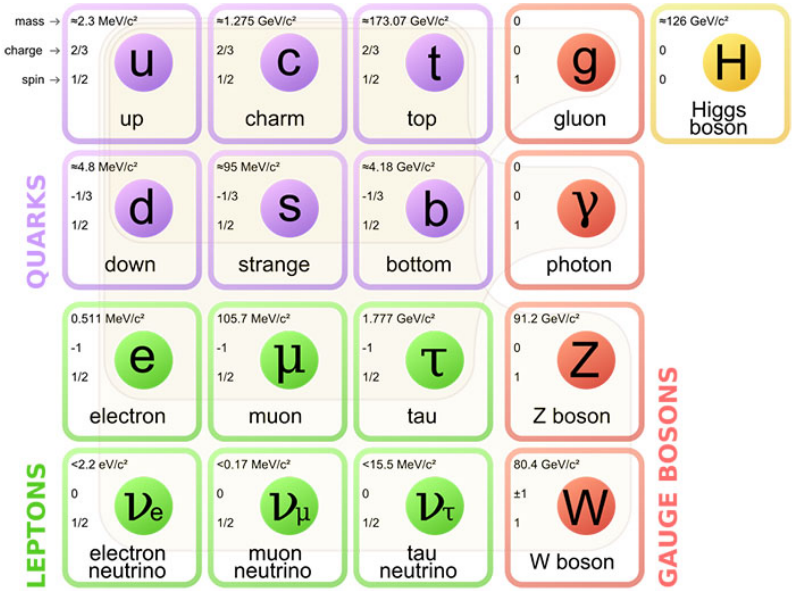
1.1 **Modelo Padrão**

No MP, os blocos fundamentais que constituem a matéria são férmions, denominados quarks e léptons e possuem spin fracionário. As interações entre as partículas são mediadas por partículas chamadas bósons que possuem spin inteiro. A interação eletromagnética é mediada pelo fóton, um bóson de spin 1 sem massa e com carga elétrica neutra, enquanto a interação fraca é mediada por três bósons de spin 1 e massivos: os

bósons $W^{\pm}$ e Z^0. A interação forte é mediada por bósons de spin 1 sem massa chamados glúons.

Os férmions são organizados em três gerações de dubletos $SU(2)_L$, ver figura 1, onde cada geração contém dois sabores de quarks com número bariônico B = 1/3 e número leptônico L = 0 e dois léptons com B = 0 e L = 1. Cada partículas possui uma antipartícula correspondente com mesma massa e números quânticos opostos.

Figura 1 - Modelo Padrão de Física de Partículas



Fonte: O autor, 2023.

Os quarks são encontrados em 6 sabores, quark up (u), down (d), charm (c), strange (s), top (t) e bottom (b) (3). A combinação de quarks formam partículas conhecidas como Hádrons, que podem ser classificados como Mésons, quando são formados por pares de quarks e antiquarks, e Bárions, quando são formados por três quarks. Mésons não possuem antipartícula correspondente, ao contrário dos Bárions. A interação fraca consegue alterar o sabor dos quarks e esse fenômeno é descrito pela matriz Cabibbo-Kobayashi-Maskawa (CKM)

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix} = \hat{V}_{CKM} \begin{pmatrix} d \\ s \\ b \end{pmatrix}. \tag{1}$$

O formalismo das interações eletrofraca e forte são desenvolvidos via simetrias de calibre que são peça fundamental no desenvolvimento de física de partículas como um todo. O MP é baseado na teoria de calibre $SU(3)_C \otimes SU(2)_L \otimes U(1)_Y$ que sofre uma quebra espontânea de simetria tal que $SU(3)_C \otimes SU(2)_L \otimes U(1)_Y \rightarrow SU(3)_C \otimes U(1)_Q$,

esse processo é responsável pela geração de massa dos bósons do setor eletrofraco e uma partícula de spin 0 chamada bóson de Higgs. A iteração forte é descrita pelo grupo de simetria $SU(3)_C$ e a interação eletrofraca é representada pelo grupo $SU(2)_L \otimes U(1)_Y$ (4).

1.2 **Matéria Escura**

A natureza da matéria escura (Dark Matter, DM) é atualmente uma das questões mais intrigantes da física fundamental. Mesmo que a questão da matéria escura tenha suas raízes na Astrofísica e na Cosmologia, ela também é ativamente procurada em experimentos de Física de Partículas, em colisores e em experimentos de detecção de matéria escura, que visam descobrir partículas de matéria escura. O principal desafio experimental é claro, a natureza elusiva da matéria escura.

Além da matéria escura, várias outras questões cosmológicas permanecem sem resposta, como a natureza da energia escura, as propriedades do período de inflação, a existência de transições de fase no Universo primitivo e a origem da assimetria de bárions no Universo. No contexto da Física de Partículas, a matéria escura pode ser composta de uma ou várias novas partículas, esperadas para ser eletricamente neutras, sem carga de cor (incolor), fracamente interagentes e estáveis. Desde que o MP falha em fornecer um candidato de matéria escura, é necessário considerar cenários além do Modelo Padrão, que podem ter uma ampla fenomenologia em colisores. Os cenários de nova física, estudados atualmente, geralmente se baseiam em novas simetrias quebradas em altas energias ou dimensões extras, e como tal, podem impactar as propriedades do Universo primitivo, seja pela presença de novas partículas no banho térmico primordial, ou através de transições de fase.

Um dos observáveis mais importantes que ligam matéria escura, Cosmologia e Física de Partículas é a densidade cosmológica da matéria escura, chamada nesse contexto de densidade remanescente. As partículas de matéria escura originam-se do Universo muito primitivo, elas estiveram em interação com o banho térmico antes de se desacoplar dele, elas então decaíram ou se aniquilaram e observamos hoje as partículas que sobreviveram até agora.

1.2.1 Cosmologia

A Cosmologia e a Astrofísica estão estreitamente ligadas à Física de Partículas. A nova física por trás desses fenômenos é amplamente estuda em experimentos atuais como o LHC, o *Large Underground Xenon* (LUX), o PandaX e o *European Laser Timing* (ELT). A cosmologia também inclui o estudo dos estágios iniciais do universo, que atualmente

não está diretamente relacionado aos experimentos de física de partículas e também o estudo da não homogeneidade no Universo através de medidas anisotrópicas da radiação cósmica de fundo e a distribuição de matéria. Nesse tópico será trabalhado conceitos-chave sobre a evolução do Universo e evidências para existência de matéria escura sob a óptica cosmológica.

1.2.1.1 Propriedades do Universo e o desvio para o vermelho

O Universo é homogêneo e isotrópico, isto é, todas as regiões do espaço são iguais e não há direção preferencial. De acordo com essas propriedades, podemos imaginar sua geometria como uma esfera tridimensional, um espaço euclidiano ou um hiperbolóide tridimensional. A métrica do espaço-tempo que comporta essas condições é chamada de Friedmann-Lemaitre-Robertson-Walker (FLRW)

$$ds^2 = -dt^2 + a^2(t)dx^2, \tag{2}$$

onde dx^2 são as coordenadas do espaço geométrico e $a(t)$ é o fator de escala dependente diretamente do tempo, esse fator comporta outra propriedade básica do Universo na métrica FLRW: o Universo se expande. As coordenadas dx^2 são comóveis, isto é, o sistema de coordenadas que acompanha a expansão do Universo, portanto, um objeto estacionário no espaço possui a mesma coordenada em todos os instantes. Então, em um Universo em expansão, o fator de escala cresce com o passar do tempo e a distância entre dois objetos com coordenadas espaciais fixas cresce. Essa formulação comporta o fenômeno das galáxias se afastando umas das outras.

A expansão do Universo acarreta o desvio para o vermelho cosmológico, ou seja, o comprimento de onda do fóton dilata. O comprimento de onda (λ_e) emitido por um objeto distante no espaço é dado por

$$\lambda = (1 + z)\lambda_e, \tag{3}$$

onde z é o desvio para o vermelho ($\frac{a(t_0)}{a(t_e)} - 1$)). Portanto, quanto mais longe uma fonte de luz está do observador, mais tempo a luz leva, medida em dado instante de tempo t_0, para percorrer a distância até o observador. Objetos que experimentam grande desvio para o vermelho estão muito distantes do observador no espaço-tempo. Esse fenômeno tem valor inestimável para observação de objetos no Universo e o estudo de suas características.

1.2.1.2 Evidências de Matéria Escura

Uma das evidências mais antigas da existência de matéria escura foi encontrada por Zwicky nos anos 30 enquanto estudava o desvio vermelho de aglomerados de galáxias. Utilizando o Teorema de Virial é possível estimar a massa total do aglomerado em função da velocidade de rotação média observada das galáxias e o raio gravitacional, estimado através das posições projetadas dos objetos. Em contraste com a massa observada no aglomerado, Zwicky percebeu ser necessário que o aglomerado possuísse muito mais massa para que as galáxias permanecessem no aglomerado (5).

Outra contribuição notável na caracterização do problema de matéria faltante são as curvas de rotação "planas" de objetos distantes em galáxias espirais, as primeiras medições foram realizadas independentemente por Albert Bosma (6) e Vera Rubin e seus colaboradores (7). Curvas de rotação são definidas através da velocidade linear ν em função da distância ao centro da galáxia r e caso a galáxia fosse um disco compacto e rígido a velocidade angular ω deveria ser constante, isto é, $\nu = \omega r$. Se todo massa estivesse concentrada na parte central luminosa da galáxia, a velocidade dos objetos distantes iria decrescer com a distância r proporcional a $r^{-1/2}$, vindo diretamente da segunda lei de Newton,

$$\nu = \sqrt{\frac{GM(r)}{r}}, \tag{4}$$

onde G é a constante gravitacional e M a massa da galáxia em função dos objetos dentro da distância r. Esse fenômeno é similar as órbitas de planetas definidas pelas leis de Kepler. No entanto, o fato descoberto por Bosma e Rubin é que as curvas de rotação galácticas são planas para grandes valores r (maiores que a extensão do disco estelar) e consequentemente a velocidade de rotação é constante (8). Portanto, como a maioria da matéria bariônica das galáxias espirais está presente na região central, é necessário a presença de uma matéria não-luminosa (escura) nos arredores da galáxia com a capacidade de influenciar gravitacionalmente a dinâmica de rotação.

Ainda seria possível argumentar que o problema de massa faltante poderia ser resolvido reformulando a teoria gravitacional, uma das alternativas à teoria newtoniana é a *Modified Newtonian Dynamics* (MOND) que obtêm sucesso ao explicar a dinâmica da curva de rotação de galáxias (9).

1.2.2 Física de Partículas

O mecanismo Brout-Englert-Higgs explica a massa não nula dos bóson do setor eletrofraco $W^{\pm}$ e Z e a massa dos férmions do Modelo Padrão através de iterações renormalizáveis e invariantes de calibre e propõe a existência de uma partícula fundamental, o bóson de Higgs (H) (10). Esse mecanismo se dá na quebra espontânea de simetria eletrofraca no campo dubleto Ψ de um escalar complexo $SU(2)_L$ com hipercarga $\frac{+1}{2}$ e um potencial escalar

$$V(\Psi) = \mu^2 \Psi^{\dagger}\Psi + \lambda(\Psi^{\dagger}\Psi)^2, \tag{5}$$

onde Ψ é o campo de Higgs, μ é o valor esperado no vácuo (VEV) e λ é o termo de autoacoplamento do campo de Higgs.

O formato do potencial escalar $V(\Psi)$ ainda precisa ser confirmado experimentalmente, contudo é um lugar matemático que se pode antecipar efeitos de uma nova física além do Modelo Padrão. O campo de Higgs pode interagir com matéria escura através de um setor além do MP e pode ser um fator crucial em explicar a assimetria bariônica no Universo (11).

Candidatos amplamente estudados para matéria escura são WIMP e Áxions. WIMPs são uma classe de candidatos para matéria escura que flutuam desde neutrinos até bósons (12). O mecanismo mais interessante da fenomenologia dos WIMPs como matéria escura em um cenário do setor de Higgs expandido é a geração de densidade cosmológica, descrito pela equação de Boltzmann sobre a hipótese de evolução cosmológica padrão do Universo (13). Após atingir o equilíbrio térmico nos primeiros estágios da evolução do Universo, a matéria escura desacopla em uma temperatura típica, chamada de *freeze-out*, aonde é possível estimar a densidade cosmológica em função da seção de choque de aniquilação de um par de matéria escura.

Áxion é um pseudobóson de Goldstone não massivo que surge da quebra espontânea da simetria de Peccei-Quinn (14), foi originalmente introduzido como uma forma de solucionar o problema de violação de carga-paridade em física de partículas. O que torna o Áxion um candidato a matéria escura é seu caráter pseudoescalar, interage gravitacionalmente e muito fracamente com oturos campos, de forma geral essa partícula faz parte de uma classe de partículas chamadas *Axion Like Particles* (ALPs). O Áxion que surge da Cromodinâmica Quântica é um modelo restrito devido sua relação fixa com sua massa e constate de decaimento, em modelos mais gerais, além do MP, um espaço de parâmetros maior é estudado ao tornar esses dois parâmetros independentes (15).

1.2.3 *Two Higgs Doublet Model* (2HDM)

Esse modelo surge da extensão do setor escalar do MP quando adicionamos um segundo campo dubleto escalar complexo $SU(2)_L$ com hipercarga $\frac{\pm 1}{2}$ (10), a equação 6, assumindo conservação CP e com uma quebra espontânea de simetria suave $\mathbb{Z}_2$, isto é, $\Psi_1 \to \Psi_1$ e $\Psi_2 \to -\Psi_2$, lê-se

$$V(\Psi_1,\Psi_2) = +m_{11}\Psi_1^\dagger\Psi_1 + m_{22}\Psi_2^\dagger\Psi_2 - [m_{12}\Psi_1^\dagger\Psi_2 + h.c.] + \frac{1}{2}\lambda_1(\Psi_1^\dagger\Psi_1)^2 + \frac{1}{2}\lambda_2(\Psi_2^\dagger\Psi_2)^2 + \lambda_3(\Psi_1^\dagger\Psi_1)(\Psi_2^\dagger\Psi_2) + \lambda_4(\Psi_1^\dagger\Psi_2)(\Psi_2^\dagger\Psi_1) + [\frac{1}{2}\lambda_5(\Psi_1^\dagger\Psi_2)^2 + h.c.]. \tag{6}$$

O espectro de partículas possui 5 bósons de Higgs (diferente de 1 no MP), onde 2 são carregados e 3 neutros. De modo que um dos bósons neutros é um pseudoescalar (CP-odd) (A_0, com massa M_{A_0}) e os outros dois são escalares (CP-even) (h e H, com massa $M_h < M_H$) (16). O modelo 2HDM possui tipos I e II descritos pela forma que os léptons e quarks se acoplam nos dubletos de Higgs. No tipo I, apenas um dubleto se acopla nos férmions, contudo, no tipo II o membro neutro de um dubleto se acopla apenas nos quarks *up* e o membro neutro do outro dubleto se acopla apenas nos quarks *down* e léptons.

A mistura dos dois Higgs CP-even é descrita pelo ângulo de rotação α e $\tan\beta$, onde $\tan\beta \equiv \nu_2/\nu_1$ (valores esperados do vácuo das componentes neutras dos campos dubletos $\Psi_{1,2}$). Os acoplamentos do Higgs para com os bósons $V = W^\pm, Z$ para os estados h e H se dá por

$$\frac{g_{hVV}}{g_{h_{SM}VV}} = \sin\beta - \alpha \tag{7}$$

$$\frac{g_{HVV}}{g_{H_{SM}VV}} = \cos\beta - \alpha. \tag{8}$$

Nos dois limites de alinhamento, $\sin\beta - \alpha \to 1$ ou $\cos\beta - \alpha \to 1$, os bósons h e H tem acoplamentos a partículas do MP iguais ao bóson de Higgs do MP, isto é, como os limites são mutuamente exclusivos apenas um dos dois bóson podem assumir as mesmas características do Higgs.

1.2.4 $\underline{\bar{b}bZ(\to \ell\bar{\ell}) + \not{E}_T(a \to \chi\bar{\chi})}$

O férmion χ (matéria escura) assume o estado de singleto sob as interações de calibre do MP se acoplando a um singleto pseudoescalar real a_0 capaz de mediar as interações entre DM e os férmions do MP através do potencial

$$V_{dark} = \frac{m_{a_0}^2}{2} a_0^2 + m_\chi \bar{\chi}\chi + y_\chi a_0 \bar{\chi}^i \gamma^5 \chi, \tag{9}$$

onde y_χ é o acoplamento de Yukawa do férmion χ. Em especial, a invariância de calibre $SU(2)_L \times U(1)_Y$ requer a existência de novos estados além do MP, de uma partícula de DM e um pseudoescalar mediador (2). Nesse contexto é introduzido o modelo 2HDM (1) como uma realização natural de uma interface entre DM e o MP renormalizável e invariante de calibre tal que o setor do Higgs é expandido com um segundo dubleto de Higgs. Porém, o portal entre o setor visível e o setor escuro ocorre através do potencial

$$V_{portal} = i\kappa a_0 H_1^\dagger H_2 + h.c., \tag{10}$$

que tem como consequência a mistura do possível estado pseudoescalar A_0 do modelo 2HDM com o pseudoescalar mediador a_0, produzindo dois pseudoescalares A e a, de massas M_A e M_a, respectivamente, definidos a partir do ângulo de mistura θ conforme

$$a = c_\theta a_0 - s_\theta A_0, \tag{11}$$

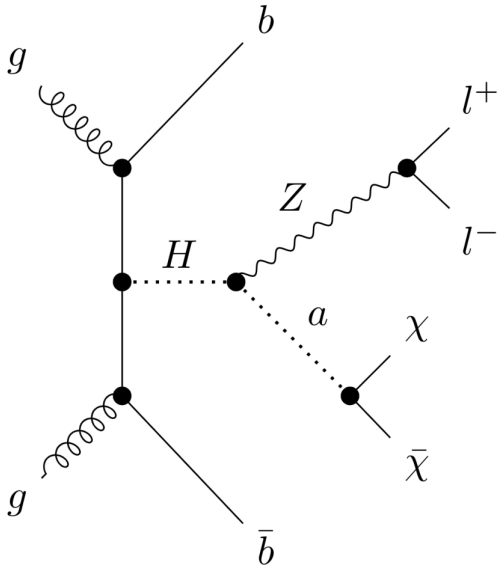
$$A = c_\theta A_0 + s_\theta a_0, \tag{12}$$

onde $c_\theta \equiv \cos\theta$ e $s_\theta \equiv \sin\theta$. Essa mistura permite que a e A se acoplem simultaneamente a DM e os férmions do MP, possibilitando um portal entre os setores visíveis e setores de matéria escura além do MP.

Essa análise se restringe ao modelo 2HDM de tipo II onde a relação entre os ângulos α e β é tal que $\beta - \alpha = \frac{\pi}{2}$ (limite do ângulo de alinhamento) e o escalar h se comporta exatamente como o bóson de Higgs do MP com massa 125 GeV. No artigo de referência (2), é considerado para sondagem $m_\chi = 45$ GeV e o diagrama de Feynman que descreve o processo de decaimento é apresentado na figura 2, onde a produção de um bóson de Higgs pesado do modelo 2HDM e dois jatos provenientes do quark bottom acontece através do mecanismo de *gluon fusion* (17), dando origem ao estado final $\bar{b}bZ(\to$

$\ell\bar{\ell}) + \not{E}_T(a \to \chi\bar{\chi})$.

Figura 2 - Processo de decaimento estudado na análise



Fonte: O autor, 2023.

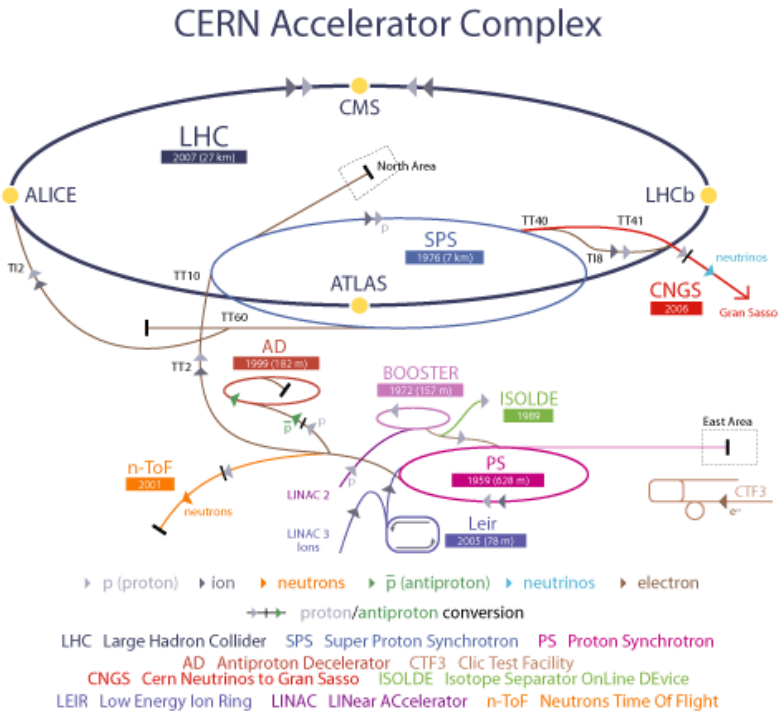
2 O LHC E O EXPERIMENTO CMS

Experimentos de Física de Altas Energias têm sido conduzidos no CERN desde 1954. O colisor LEP (*Large Electron Positron*) foi um dos mais notáveis, operando de 1989 a 2000 e ajudando a estabelecer a validade do Modelo Padrão. Na sua primeira fase, ele colidiu elétrons e pósitrons para produzir bósons Z^0 e, na segunda fase, produziu os bósons $W^{\pm}$. Além da produção do bósons do setor eletrofraco do SM, outro feito histórico realizado no LEP foi a determinação das constantes de acoplamento do SM com uma precisão sem precedentes (18) e predições para as massas do quark-top e bóson de Higgs.

Para explorar ainda mais a estrutura da matéria, foi criado o LHC (*Large Hadron Collider*), sendo atualmente o maior e mais poderoso acelerador de partículas do mundo. O LHC está localizado entre a Suíça e a França, a cerca de 100 metros abaixo da superfície e é um túnel circular de 27 km feito de magnetos supercondutores. O LHC é um colisor de hádrons, primariamente feixes de prótons, esse tipo de colisor permite a exploração de ressonâncias com valores de massas mais altos, o que culminou na descoberta do bóson de Higgs em 2012.

O complexo de aceleradores, ver figura 3, é usado para aumentar a energia do feixe de partículas antes que ele entre no túnel principal para colisão (19). No LHC, quatro experimentos são realizados por colaborações internacionais, cada um com seu próprio detector de partículas único localizado em ou perto de quatro pontos de colisão. Os dois maiores experimentos, *A Toroidal LHC ApparatuS* (ATLAS) e *Compact Muon Solenoid* (CMS), são detectores de propósito geral projetados para investigar uma ampla gama de fenômenos físicos, enquanto *A Large Ion Collider Experiment* (ALICE) e *LHC-beauty* (LHCb) são especializados em analisar colisões relacionadas a fenômenos específicos. Prótons são obtidos removendo elétrons dos átomos de hidrogênio sendo submetidos a uma sequência de aceleradores antes de serem inseridos no LHC. Campos magnéticos intensos são usados em todos os aceleradores circulares para curvar a trajetória dos prótons durante a aceleração. Toda a infraestrutura é suportada por um complexo sistema de resfriamento e alimentação de energia, bem como um sistema de controle e aquisição de dados.

Figura 3 - Complexo de Aceleradores do CERN



Legenda: Ilustração do complexo de aceleradores e seus sub-detectores.
Fonte: European Organization for Nuclear Research, 2023.

A luminosidade é uma quantidade importante relacionada ao acelerador, definida como a razão de eventos nas colisões por unidades de área

$$L = F\frac{\gamma f k_b N_p^2}{4\pi\epsilon_n\beta^*}, \tag{13}$$

onde F é o valor de redução geométrico relacionado ao ângulo entre os feixes no cruzamento, β^* é o valor belatron (função de amplitude que mensura o quanto os ímãs conseguem focar o feixe no ponto de interação), ϵ_n é a emitância transversa normalizada (mensura a divergência e a compactação das nuvens resultante de efeitos dos feixes), N_p é o número de partículas por pacote, f é a frequência de revolução do LHC e γ é o fator relativístico. Ao decorrer da tomada de dados as propriedades do feixe são degradas, implicando em uma diminuição da luminosidade instantânea, então, a luminosidade integrada ($\mathcal{L}_{int}$) em um instante de tempo t é proporcional ao tempo τ de vida útil da luminosidade (21) devido à degradação do feixe e torna-se uma quantidade melhor a ser analisada, sendo definida por

$$\mathcal{L}_{int} = L\tau(1 - e^{-\frac{t}{\tau}}), \tag{14}$$

O número de eventos (N) produzido na colisão é calculado através do produto entre a luminosidade integrada e a seção de choque (σ_{tot}), que por sua vez é a probabilidade de ocorrência de uma interação. A seção de choque total do evento é calculada através do somatório de seções de choque parciais relacionadas com ocorrência de cada estado final possível no decaimento de uma partícula (4), portanto

$$\sigma_{tot} = \sum_{i=1}^{n} \sigma_i, \tag{15}$$

$$N = \mathcal{L}_{int}\sigma_{tot}. \tag{16}$$

2.1 **O detector CMS**

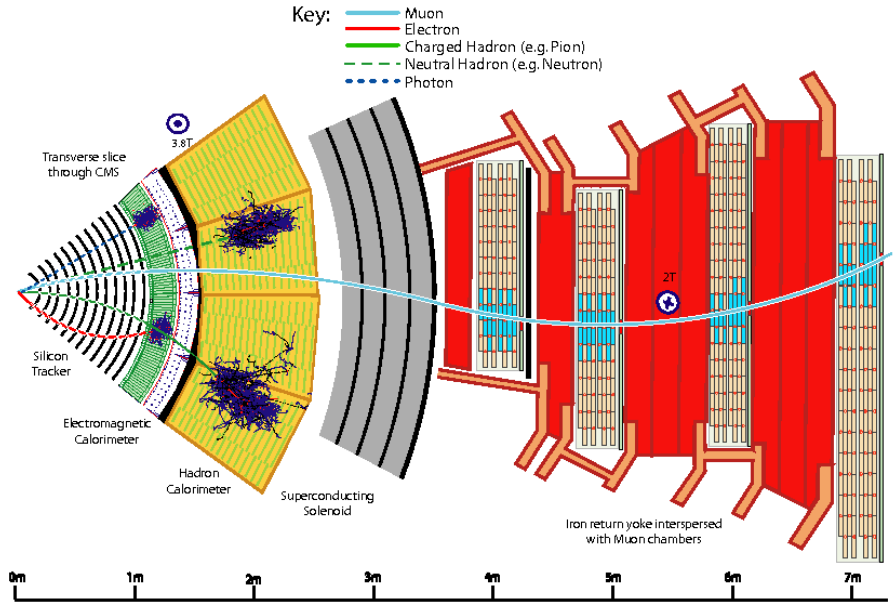
O Compact Muon Solenoid é um detector de partículas localizado no ponto de interação 5 do LHC no CERN, que emprega um solenoide supercondutor para gerar um campo magnético para desviar as trajetórias de partículas carregadas, permitindo sua identificação e medição de suas propriedades. O detector CMS, ver figura 4 foi especificamente projetado para detectar e estudar as propriedades do bóson de Higgs e física além do Modelo Padrão. O solenoide supercondutor do detector CMS é um dos maiores e mais poderosos ímãs supercondutores já construídos, possuindo aproximadamente 13 metros de largura e 6 metros de diâmetro interno oferecendo um campo magnético de 3,8 T (22).

O sistema de coordenadas, ver figura 5, tem sua origem no ponto de colisão das partículas (24), a coordenada z está alinhada com a direção do feixe e paralelo ao campo magnético do CMS, a coordenada y é orientada para cima e a coordenada x aponta para o centro do anel do LHC. Em coordenadas polares, o ângulo polar (θ) é definido a partir do eixo z e o ângulo azimutal (ϕ) é definido com relação ao plano $x - y$. Ainda é possível definir a pseudorapidez (η) como

$$\eta = -\ln\left[\tan\left(\frac{\theta}{2}\right)\right], \tag{17}$$

onde θ é o ângulo polar. Essa quantidade é utilizada para descrever os ângulos cujas partículas são detectadas no detector e as propriedades dessas partículas.

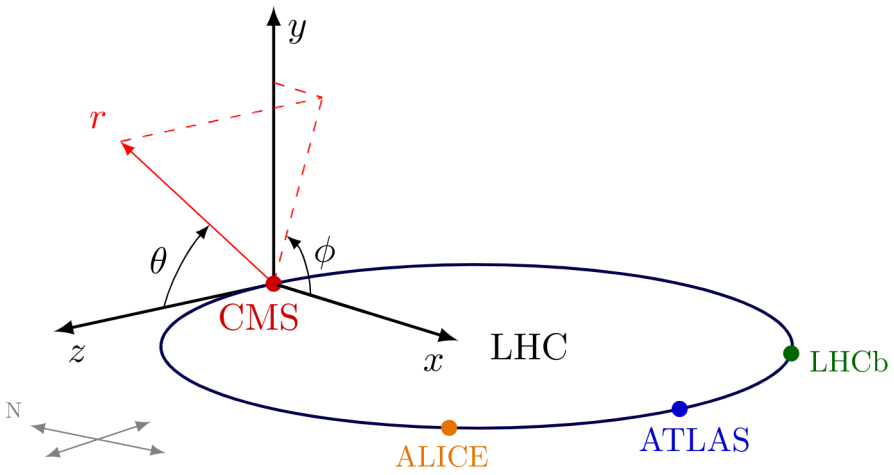
Figura 4 - Visão esquemática do detector do CMS



Legenda: Sub-detectores do CMS e como partículas interagem com eles.

Fonte: SIRUNYAN, 2017, p. 2.

Figura 5 - Esquema do sistema de coordenadas do CMS



Legenda: Sistema de coordenadas do CMS com base nas direções cardeais e as posições dos experimentos no LHC.

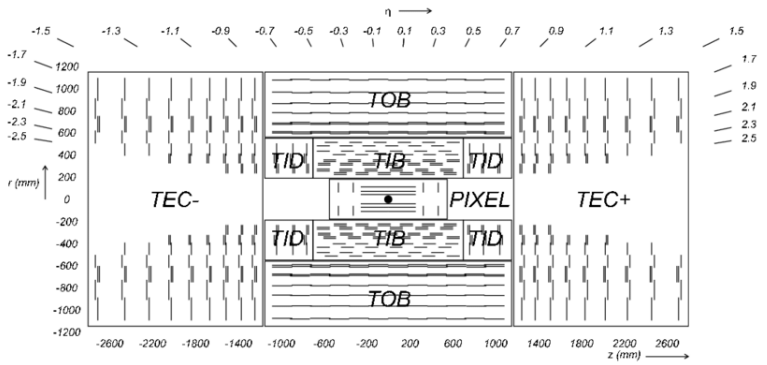
Fonte: NEUTELINGS, 2023.

2.2 **Sistema de Trajetografia**

O sistema de trajetografia (ou tracking system), ver figura 6, é responsável por mensurar com alta precisão as trajetórias das partículas carregadas produzidas nas colisões dos prótons. A mudança na trajetória provocada pelo campo magnético intenso permite a identificação da partícula, além disso, é possível determinar a carga elétrica e mensurar o momentum. A eficiência na reconstrução precisa dessas quantidades é maior na região delimitada por $|\eta| < 2,5$.

A distância com relação ao ponto de interação é pequena o suficiente para que o sistema fosse construído com uma alta granularidade, permitindo a distinção da origem da partícula quando existe mais de uma interação por *bunch*. A estrutura é construída utilizando detectores de tiras de silício e detectores de pixel, organizados paralela e ortogonalmente ao feixe, são denominadas respectivamente de barris e *endcaps*.

Figura 6 - Sistema de trajetografia do CMS



Legenda: Visão esquemática do sistema de trajetografia do CMS.
Fonte: BAUER, 2010, p. 16.

Os detectores de pixel possuem cerca de 66 milhões de sensores com 100 $\mu m \times$ 150 μm e estão alocados em 3 três camadas de barris e 2 camadas de *endcaps* localizados a 34 cm e 46,5 cm do ponto de interação, respectivamente. Devido à proximidade com o ponto de interação, esse detector fornece a determinação precisa do parâmetro de impacto da colisão e a reconstrução dos vértices primários e secundários dos eventos.

O detector de tiras, por outro lado, pode ser divido nas partes interior e exterior. A parte interior é formado pelo *Tracker Inner Barrel* (TIB) com 4 camadas de fitas e pelo *Tracker Inner Disks* com 4 camadas de discos de fitas na extremidade do barril. A parte externa é formada pelo *Tracker Outer Barrel* com 6 camadas de barril e pelo *Tracker End-Cap* com 9 camadas em forma de discos na extremidade de cada barril.

2.3 **Sistema de Calorimetria**

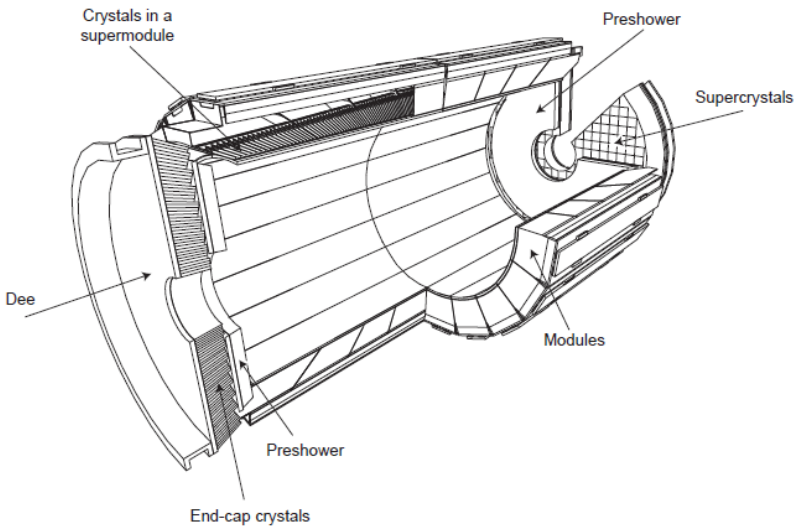
Calorímetros desempenham a função da absorção e amostragem de energia das partículas incidentes e diferente do sistema de trajetografia são sensíveis a partículas neutras (27). O CMS possui dois tipos de colorímetros: o calorímetro eletromagnético (ECAL) e o calorímetro hadrônico (HCAL).

2.3.1 Calorímetro Eletromagnético (ECAL)

O ECAL, ver figura 7, é um calorímetro hermético e homogêneo constituído por 75848 cristais cintilantes de tungstato de chumbo (PbWO4) usados para medir a energia, por absorção, de partículas que interagem eletromagneticamente (elétrons, fótons, hádrons carregados) (23). O sinal elétrico é capturado através da luz coletada por foto-detectores, localizados no final de cada cristal, quando os cristais emitem luz que é proporcional a energia da partícula (28).

É composto na região central por um barril (EB) organizado em 35 super módulos, cada contendo 1700 cristais e fechado por duas tampas (*endcaps*) com 7324 cristais cada. Detectores de silício *preshower* (ES) estão instalados na parte frontal dos endcaps do calorímetro, com objetivo de identificar e discriminar os dois fótons (de baixa energia) produto do decaimento de píons neutros dos fótons primários (de alta energia). O barril cobre uma região de $|\eta| < 1{,}48$ que se estende até $|\eta| < 3{,}0$ com os *endcaps*. Com os detectores *preshower* é possível cobrir uma região $1{,}65 < |\eta| < 2{,}6$.

Figura 7 - Calorímetro Eletromagnético (ECAL) do CMS



Legenda: Visão esquemática do calorímetro eletromagnético do CMS.
Fonte: BARTOLONI, 2013, p. 2.

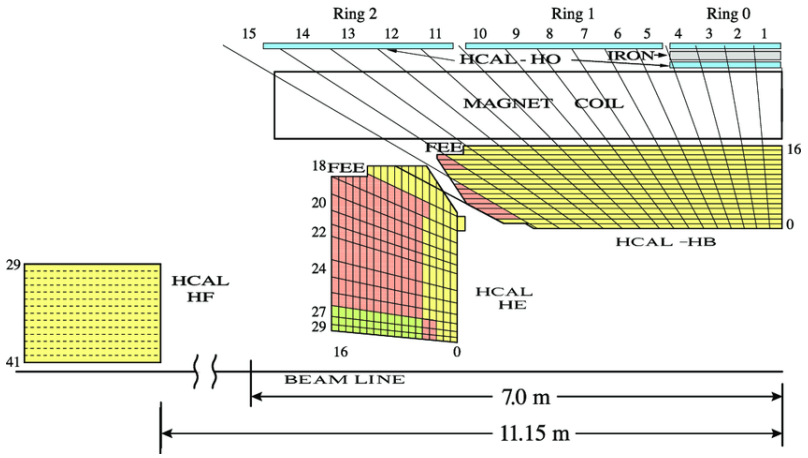
2.3.2 Calorímetro Hadrônico (HCAL)

O HCAL, ver figura 8, é um calorímetro por amostragem hermética (30) que possibilita mensurar a energia depositada por hádrons (prótons, nêutrons, píons e káons) produzidos nas colisões, além de ajudar na identificação de fótons e léptons. Sua estrutura é construída utilizando cintiladores plásticos e camadas de aço e bronze, de tal forma que quando um hádron penetra no sub-detector ele interage com os núcleos pesados e produzindo um chuveiro (*jets*) que emitem luz quando atravessam os cintiladores.

Além disso, outra característica diferencial do HCAL é a capacidade de medir indiretamente partículas não-interagentes e sem carga, como os neutrinos, a quantidade medida é denominada energia transversa perdida (ou *Missing Transverse Energy*, MET). Essa quantidade é determinada através da conservação de energia e momentum do decaimento das partículas, ou seja, falta de energia para satisfazer a condição de conservação indica, indiretamente, a produção de partículas indetectáveis pelo calorímetro. Portanto, o desbalanço de energia e momentum é extremamente útil em análises que envolvem neutrinos ou partículas propostas em modelos além do modelo padrão.

Esse calorímetro cobre uma região de pseudorapidez $|\eta| < 5{,}2$, sendo divido em quatro partes: *Hadron Calorimeter Barrel* (HB), *Hadron Calorimeter Endcapds* (HE), *Outer Hadron Colorimeter* (HO) e *Forward Hadron Calorimeter* (HF). O HB cobre a região $|\eta| < 1{,}3$. A região $1{,}3 < |\eta| < 3$ compreende 34% das partículas no estado final sendo abrangida pelo HE (31). O HO compreende a região de $|\eta| < 3$ e é responsável por identificar os chuveiros que acontecem após o HB, bem como medir a energia depositada pelos mesmos. Por fim, o HF está situado nas extremidades do CMS, posicionado a 11,2 metros do ponto de interação (32), cobre a região $3 < |\eta| < 5{,}2$ e busca detectar a luz produzida a partir do efeito de Cherenkov.

Figura 8 - Calorímetro Hadrônico (HCAL) do CMS



Legenda: Visão esquemática do calorímetro hadrônico do CMS.

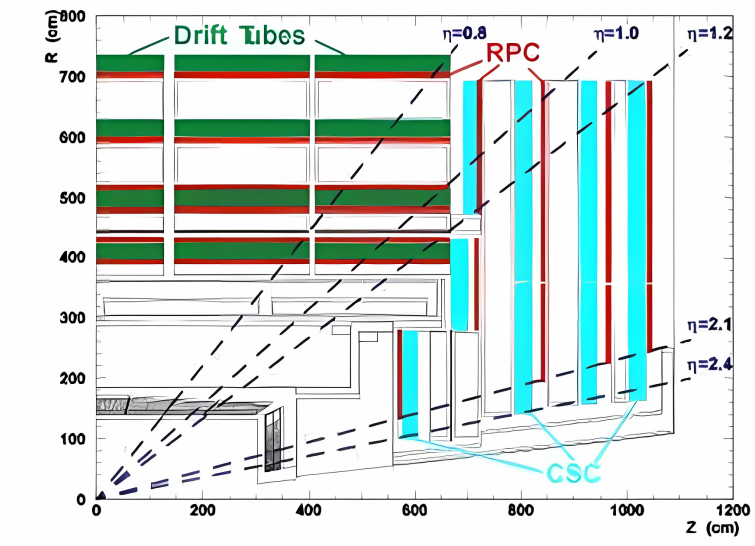
Fonte: CMS COLLABORATION, 2010, p. 3.

2.4 **Câmaras de Múons**

As câmaras de múons (ou *Muon System*), ver figura 9 localizadas na região mais externa do experimento CMS, utilizam três tipos diferentes de tecnologias de detectores de ionização de gás para medir a trajetória e o momento dos múons. Essas tecnologias incluem Drift Tubes (DTs), Cathode Strip Chambers (CSCs) e Resistive Plate Chambers (RPCs) (34). A necessidade de construir um módulo específico para detecção de múons do sub-detector é devido a sua característica fracamente interagente, conseguem atravessar os calorímetros do CMS deixando um sinal fraco e também atravessam as camadas do *yoke*.

Os DTs, que possuem uma precisão de 250 μm na medida espacial, são divididos em 240 câmaras com células de deriva. Eles são colocados na região do barril, onde a ocupação e o ruído de fundo são baixos, e o campo magnético residual está principalmente contido no *yoke* de retorno. Os CSCs, que fornecem uma medida precisa da direção ϕ e uma resolução espacial de cerca de 200 μm, são compostos por 540 câmaras trapezoidais localizadas nas extremidades do CMS. Os RPCs são câmaras de espaço duplo operadas em modo avalanche e fornecem uma resposta rápida no tempo para a atribuição de cruzamento de pacote não ambígua, têm uma resolução espacial mais grossa. Eles estão localizados nas regiões do barril e das extremidades, cobrindo uma região de $|\eta| < 1{,}61$, com uma resolução espacial de (0,8 - 1,2) cm (35).

Figura 9 - Câmara de Múons do CMS



Legenda: Visão esquemática da Câmara de Múons do CMS.
Fonte: CHAUHAN, 2009, p. 74.

Os múons reconstruídos podem ser classificados em três tipos dependendo dos sub-detectores utilizados na reconstrução: *globalmuon* são aqueles reconstruídos utilizando

o sistema de trajetografia e *hits* nas câmaras de múons, *standalone* são reconstruídos utilizando apenas o sistema de múons; e *trackermuons* utilizam o sistema de trajetografia, uma validação pelo sistema de calorimetria e parte do sistema múons. Outra característica importante é a discriminação dos múons produzidos na colisão dos múons provenientes de raios cósmicos.

2.5 **Sistema de *Triggers***

A quantidade de informação (sinal eletrônico, sendo posteriormente reconstruído) produzido durante as colisões é altíssima, contudo, a quantidade de eventos físicos interessantes é muito baixo, dessa forma, fez-se necessário desenvolver um sistema de *triggers* para filtrar e armazenar somente eventos produzidos de interesse. O sistema é desenvolvimento utilizando primeiro um módulo de hardware chamado de *Level-1 trigger* (L1) e posteriormente um software chamado *High Level Trigger* (HLT).

O *trigger* L1 coleta informações do sistema de calorímetros e do sistema de múons e consegue reduzir a razão de eventos durante a tomada de dados de 40 MHz para 100 kHz (22), gerando uma economia de aproximadamente 39 Tb/s de dados para serem processados e armazenados. Também é chamado de *online trigger*, pois, os eventos são selecionados durante a tomada de dados.

Por outro lado, o HLT utiliza informação de todos os sub-detectores e consegue reduzir a razão de eventos de 100 kHz (*input* do L1 *trigger*) para 1 kHz. Por se tratar de um *software* muitos algoritmos são executados e utilizam critérios de seleção bem definidos para selecionar apenas eventos de interesse. Uma situação comum é alta produção em alguns processos de decaimento, então, uma forma de controlar a razão de eventos selecionados é utilizando a estratégia de *prescaling*, isto é, configuramos um *trigger* HLT com uma pré-escala definida de N, dessa forma, será armazenado apenas uma quantidade 1/N dos eventos que originalmente passaram pelo critério de seleção desse *trigger* (37). A estrutura do software é dividida em camadas, ou seja, primeiros são aplicadas algoritmos de seleção considerando apenas informação dos calorímetros e sistema de múon (*Level-2*, depois informação dos *pixels* (*Level-2.5*) e por fim o sistema de trajetografia (*Level-3*).

2.6 **Reconstrução dos objetos físicos**

Os objetos físicos utilizados nas análises de dados são produto da aplicação de diversos algoritmos que tem como *input* sinais eletrônicos produzidos nos sub-detectores e armazenados após o L1 e HLT *triggers*, dizemos que esses objetos são reconstruídos e identificados. No CMS é utilizado o algoritmo *Particle Flow* (PF) para reconstrução desses objetos.

2.6.1 *Particle Flow*

O *Particle Flow* (PF) utiliza os elementos depositados em cada sub-detector, que são combinados para identificação de partículas diferentes. O algoritmo correlaciona elementos básicos de todas as camadas do detector (*tracks* e *clusters*) para identificar cada partícula no estado final e combinar as medições correspondentes para reconstruir as propriedades das partículas. Partículas diferentes são reconstruídas usando sinais de diferentes regiões do detector, como hádrons carregados sendo identificados por uma conexão geométrica no plano $\eta - \phi$ e ausência de sinal nos detectores de múons, fótons e hádrons neutros sendo identificados pelo sinal depositado no ECAL e HCAL e *tracks* descorrelacionados, elétrons sendo identificados pela presença *tracks* e um sinal no ECAL e múons sendo identificados por *tracks* na região interna do detector conectada a outra nos detectores de múons.

2.6.2 Múons

O processo de reconstrução de múons envolve três etapas usando informações tanto das Câmaras de Múons quanto do Sistema de Trajetografia. Na primeira etapa, os *hits* do RPC e os segmentos do DT e CSC são reconstruídos dentro de cada câmara e os traços nas Câmaras de Múons são reconstruídas combinando *hits* e segmentos de diferentes seções.

Em seguida, a trajetória do múon é reconstruída combinando informações do Sistema de Trajetografia e das Câmaras de Múons. O algoritmo *Global* Muon reconstrói a trajetória do múon procurando por pares compatíveis de trajetórias no Sistema de Trajetografia e usando uma técnica de filtro de Kalman (23) para ajuste. O algoritmo *Tracker* Muons reconstrói a trajetória do múon extrapolando as trajetórias do Sistema de Trajetografia para as Câmaras de Múons e associando segmentos compatíveis. Em seguida, é realizado um ajuste completo de trajetórias para todas as combinações compatíveis, sendo selecionada a melhor correspondência.

Os critérios de seleção são ajustados com base nas propriedades dos atributos do múon e informações de outros sub-detectores. Os múons são reconstruídos se eles têm no mínimo $p_T > 0,5$ GeV/c e podem ser extrapolados para um agrupamento no sistema de múons para garantir uma boa precisão na reconstrução do objeto físico. A última etapa é a identificação de múons, cujo objetivo é suprimir múons produzidos em jatos hadrônicos que possivelmente penetraram o sistema de múons.

2.6.3 Elétrons

A reconstrução da energia depositada por elétrons, fótons e hádrons é realizada por métodos diferentes. Elétrons e fótons depositam a maioria de sua energia no ECAL, enquanto hádrons depositam a maioria de sua energia no HCAL (38). À medida que elétrons ou fótons se propagam através do material à frente do ECAL, os elétrons podem interagir e emitir fótons de bremsstrahlung ($e^- \rightarrow e^-\gamma$), a consequente perda de energia pela radiação desses fótons produz uma trajetória mais curvada devido ao campo magnético intenso e os fótons se convertem em um par elétron-pósitron (e^-e^+) deixando um sinal característico de duas trajetórias eletricamente carregadas com o vértice comum. Um algoritmo dedicado é usado para combinar os clusters dessas partículas em um único objeto para recuperar a energia primária.

2.7 **Identificação de jatos provenientes do quark bottom (*b-tagging*)**

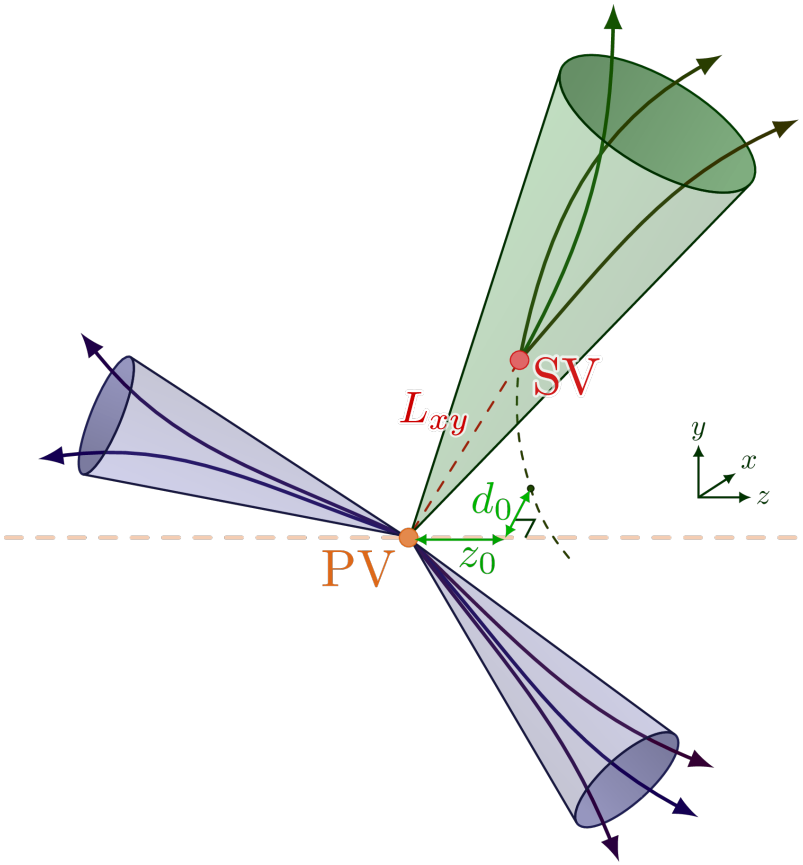
O *b-tagging* é uma ferramenta essencial para estudar processos físicos com a presença de jatos de quark *bottom* no estado final, como os presentes em setores de Higgs ($H \rightarrow bb$) e a física do quark *top* ($t \rightarrow Wb$). Ou seja, é utilizando técnicas de *b-tagging* que físicos de partículas experimentais são capaz de discriminar jatos provenientes de outros quarks de jatos provenientes de quark *bottom* e então prosseguir com a análise de dados adequada para o caso de estudo, como o caso dessa dissertação que os eventos de interesse podem ser discriminandos em grande parte através da identificação de jatos do quark *bottom*. Algumas propriedades básicas desses jatos são:

- b-quarks se fragmentam em B-hádrons;
- presença de um vértice secundário (ou *secondary vertex*, SV), deslocado do vértice primário (ou *primary vertex*, PV) devido à grande vida média, ver figura 10;
- presença de múons em jatos de quark *bottom* no decaimento de hádrons pesados.

Vértices são pontos de interação (IP) entre partículas durante colisões e de-

caimentos de partículas com vida média significativa. Em uma colisão, o IP nominal é a posição pretendida de colisão no detector e o IP real é a posição onde as partículas realmente colidem. O IP real é o vértice primário do evento. Graças a excelente resolução do sistema de trajetografia do CMS, é possível reconstruir diretamente o vértice secundário do evento, isto é, o ponto que o B-hádrons decaem. Para realizar essa tarefa, durante a primeira tomada de dados (*Run1*) foi desenvolvido o algoritimo *Adaptive Vertex Reconstruction* (AVR) para a segunda tomada de dados (*Run2*) o *Inclusive Vertex Finding* (IVF). O AVR interativamente ajusta grupos de traços como jatos ($\Delta R < 0,3$) após seleção básica e o IVF se limita ao agrupamento de *tracks* reconstruídos com $p_T > 0,8$ GeV e o ajuste do SV após arbitração de traços.

Figura 10 - Representação da produção um jato proveniente do quark *bottom*



Legenda: O vértice secundário (descrito na imagem como SV) é fator determinante na identificação do sabor do jato.

Fonte: O autor, 2023.

Portanto, os algoritmos de identificação no CMS são construídos considerando a vida média, grande massa e fração de momentum dos B-hádrons e a presença de *soft-leptons* produzidos em jatos de quark bottom. O algoritmo *Jet Probability* (JB) explora a verossimilhança da distribuição densidade de probabilidade das trajetórias que não envolvem b/c-jets através do parâmetro de impacto, sendo calculada a probabilidade de

uma trajetória ser originada do vértice primário (PV), então, é combinada a probabilidade de todas as trajetórias e designada a probabilidade do jato ser originado do PV (39). O *JetB Probability* (JBP) é similar ao JB, porém, da maior peso as quatro trajetórias mais deslocadas do evento. Outro algoritmo de baixo nível é baseado na identificação de *soft-leptons* (SL), os *taggers* SL exploram as propriedades dos múons e elétrons produzidos no decaimento semi-leptônico no decaimento de quark bottom, devido à grande massa do quark bottom, o momentum transverso do múon com relação ao eixo do jato é maior para múons originados do decaimento de B-hádrons do que para múons originados de jatos leves (udsg).

Algoritmos baseados na presença de SV, como *Combined Secondary Vertex* (CSV) e *Combined Secondary Vertex Version 2* (CSVv2) utilizam a informação da reconstrução de pelo menos um vértice secundário e a massa das partículas carregadas reconstruída no SV usada para mensurar a pureza da amostra de jatos identificados como provenientes de quark bottom (*b-tagged*) (40). Além disso, esses algoritmos utilizam modelos de aprendizado de máquina para combinação eficiente das trajetórias deslocadas com relação ao SV, *Boosted Decision Trees* (BDT) no caso do CSV e *Multi Layer Perceptron* (MLP) no caso do CSVv2. O *Conditional Mean Value Analysis* (cMVA) é outro algoritmo baseado em aprendizado de máquina que combina os discriminadores de baixo nível para produção de um discriminante mais robusto.

Por fim, com os recentes avanços no campo de redes neurais profundas (*Deep Neural Networks*, DNN), com o intuito de aprimorar a identificação de jatos provenientes do quark bottom, foram desenvolvidos os modelos DeepCSV e DeepJet (ou DeepFlavour). O DeepCSV (41) utiliza as mesmas *features* do CSVv2, contudo, considera trajetórias de partículas carregadas. O DeepJet utiliza uma arquitetura de rede neural mais complexa, incluindo nas camadas profundas redes convolucionais e recorrentes, além de possuir uma eficiência superior a todos os outros modelos atuais (42).

Variando os cortes no discriminador, obtemos diferentes eficiências dos *taggers*, aonde é possível estabelecer pontos de operação (*Working Points*, WP) como *loose* (L), *medium* (M) e *tight* (T), aonde a taxa de má identificação dos jatos estimada a partir do Monte Carlo é de 10%, 1% e 0,1%, respectivamente, para o momentum transverso de um jato de cerca de 80 GeV (39). A performance dos *taggers* não é prejudicada pela presença de *pileup* (efeitos de interações adicionais provenientes de outros vértices de interação) nos eventos, devido à boa seleção de trajetórias.

3 APRENDIZADO DE MÁQUINA

Aprendizado de máquina é uma abordagem que visa a construção de algoritmos capazes de aprender com dados e, assim, resolver tarefas de classificação e regressão. Essa abordagem pode ser classificada em três categorias principais: aprendizado supervisionado, não supervisionado e por reforço. O aprendizado supervisionado é utilizado para prever valores futuros a partir de dados rotulados. Já o aprendizado não supervisionado é aplicado na busca de padrões em dados não rotulados. Por fim, o aprendizado por reforço é usado para a tomada de decisões em ambientes dinâmicos.

Nesse trabalho serão abordados modelos de aprendizado de máquina supervisionados, aonde dado um parâmetro de entrada x_i com classe y_i é produzido um discriminante capaz de classificar novos parâmetros de entrada em classes $\hat{y}_i$. Classe é um termo utilizado para se referir a classificação de conjunto de parâmetros de entrada, isto é, em um trabalho de classificação um vetor com características que representam o sinal tem classe de valor 0 e para o *background* classe de valor 1, esses valores são arbitrários e são definidos na etapa de pré-processamento do conjunto de treino e teste.

Um conjunto de dados (ou *dataset*) é utilizado para produção do modelo, em tarefas supervisionadas dividimos o conjunto de dados em conjuntos de treino e teste que são utilizados, respectivamente, nas etapas de treinamento (processo de descoberta de padrões) e validação (processo de revisão da performance do discriminador). A etapa de descoberta de padrões é um processo matemático de extração de informação de variáveis (ou *features*) através de modelos matemáticos particulares para cada tipo modelo, contudo, em geral é definido uma função de perda (ou *loss function*) dedicada a mensurar os erros cometidos pelo modelo na etapa de aprendizado. A otimização da função de perda, através de algoritmos como gradiente descente, é um processo crucial para o cálculo dos melhores pesos definidos pelo modelo no qual aplicados sobre as *features* geram uma tomada de decisão.

3.1 Gradiente descendente

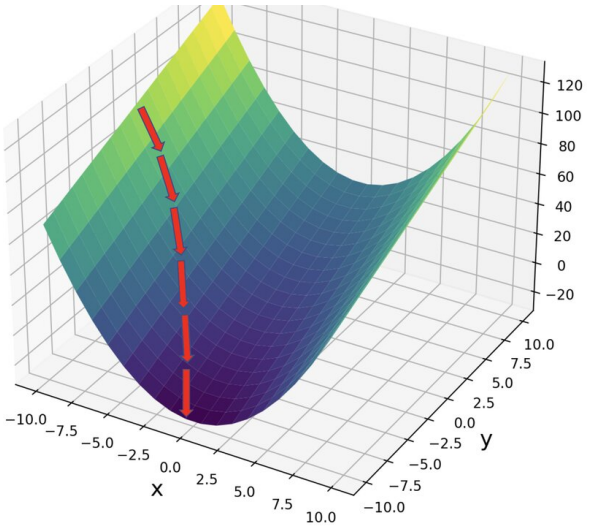
O gradiente descendente (*gradient descent*) é um algoritmo de otimização usado para minimizar uma função de perda (*loss function*). Para isso, os parâmetros de entrada são atualizados iterativamente na direção negativa do gradiente da função de perda (direção de queda mais acentuada). O processo de atualização só é finalizado quando a mudança na função de perda é tão pequena quanto se queira ou um número de iterações máximo é alcançado. A figura 11 ilustra possíveis caminhos para mínimos de uma função através do gradiente descente. A cada iteração os parâmetros de entrada são

atualizados conforme

$$x(i+1) = x(i) - \eta \nabla f(x(i)), \tag{18}$$

onde $x(i)$ é os parâmetros de entrada atuais, f é a função de perda, $\nabla f(x(i))$ é o gradiente da função de perda com respeito a $x(i)$ e η é o *learning rate* (43). O *learning rate* controla o tamanho do passo tomado na direção do gradiente, ou seja, valores pequenos para η irão resultar em um tempo maior de convergência, porém, uma maior probabilidade de convergir para o mínimo global, enquanto valores grandes de η convergirão mais rápido, mas, podem estimar incorretamente o mínimo global ou até divergir.

Figura 11 - Ilustração do algoritmo gradiente descentente



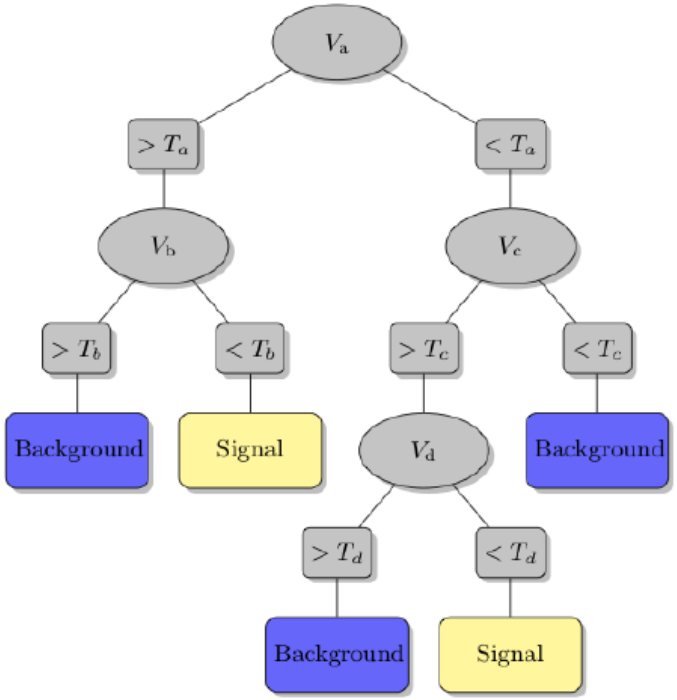
Fonte: O autor, 2023.

3.2 Árvores de decisão

Árvores de decisão são uma ferramenta popular e poderosa no campo do aprendizado de máquina, elas são um tipo de algoritmo que pode ser usado tanto para tarefas de classificação quanto de regressão (44). As árvores de decisão funcionam dividindo recursivamente os dados em subconjuntos com base nos valores das características de entrada. O resultado é um modelo em forma de árvore que pode ser usado para fazer previsões sobre novos dados. A figura 12 exemplifica um modelo baseado em árvores de decisão.

No processo de aprendizado de árvores de decisão o objetivo é encontrar a melhor separação dos dados em cada nó de modo que a árvore resultando tenha a melhor classificação ou acurácia em uma predição. Para isso, uma função de custo (*cost function*)

Figura 12 - Ilustração de um modelo de árvore de decisão



Legenda: Visão esquemática de uma árvore de decisão utilizada para discriminação de classes *background* e *signal* na análise de Ondas Gravitacionais, onde cada variável V_i é comparada com um limiar T_i para tomada de decisão do sistema.

Fonte: ADAMS ET AL, 2013, p. 3.

é utilizada para medir a qualidade da separação, então, a função de custo é minimizada para encontrar a melhor separação dos dados (46).

A *Gini impurity* é uma função de custo muito utilizada no aprendizado de uma árvore de decisão sendo definida por

$$Gini(S) = 1 - \sum p(i)^2, \tag{19}$$

onde S é um conjunto de dados e $p(i)$ é a probabilidade de um item i no conjunto pertencer a uma determinada classe.

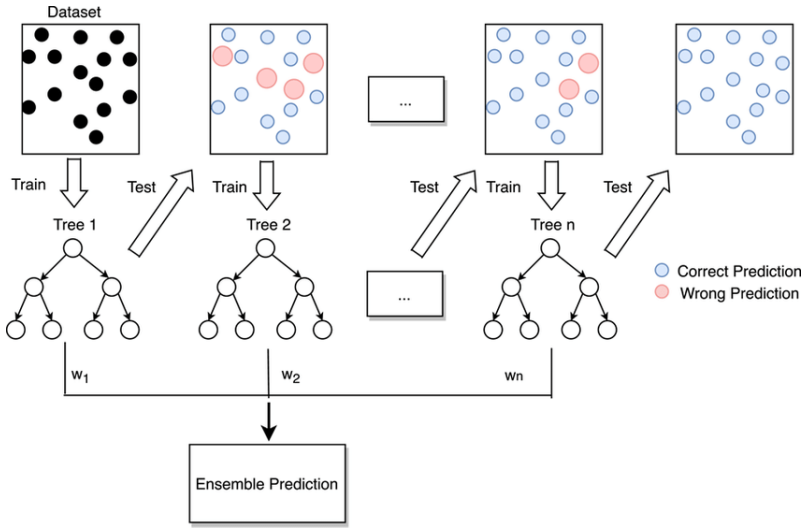
Uma das principais vantagens das árvores de decisão é que elas são fáceis de entender e interpretar, pois, a estrutura em árvore fornece uma representação clara das decisões que foram tomadas.

3.3 **XGBoost**

O XGBoost (XGB) (47) é uma biblioteca de código aberto que oferece uma implementação eficiente do algoritmo de *gradient boosting* (48). Ele é particularmente popular em competições do Kaggle e já foi usado para vencer muitas soluções, inclusive no ramo de Física de Altas Energias utilizando dados do ATLAS (49). As principais características do XGB incluem processamento paralelo, regularização, *early-stopping* e tratamento de valores ausentes, oferecendo uma ampla variedade de métricas de avaliação para classificação e regressão.

A implementação do *gradient boosting* funciona ajustando uma sequência de árvores de decisões fracas (*weak learners*), onde cada modelo tem o objetivo de corrigir os erros do modelo anterior na sequência, ver figura 13. Isso é feito minimizando uma função de perda, como o erro quadrático médio para regressão ou a perda logarítmica para classificação, usando o gradiente descendente.

Figura 13 - Ilustração da sequência de treino do XGB



Legenda: Diagrama do método de treinamento de *gradient boosting*, o classificador final (*ensemble*) é a composição de múltiplos classificadores mais fracos, de modo que cada modelo posterior aprende com os erros do modelo anterior.

Fonte: ZHANG ET AL, 2021, p. 6.

Iterativamente, o XGB calcula o gradiente da função de perda com respeito as predições da árvore e então calcula o hessiano da mesma forma (47). A derivada de segunda ordem da função de perda, calculada a partir do hessiano, melhora a eficiência e acurácia do algoritmo de *gradient boosting*. A matriz hessiana, definida por

$$H_{ij} = \frac{\partial^2 f(y_i, \hat{y}_i)}{\partial \hat{y}_i \partial \hat{y}_j} \tag{20}$$

onde, f é a função de perda, y_i é a classe verdadeira do dado de entrada, $\hat{y_i}$ é a classe predita. Durante a etapa de minimização da função de perda, a atualização da direção ótima d_i de cada instância de treinamento é calculada resolve a equação quadrática

$$d_i = -\frac{\nabla f(x(i))}{H_i + \lambda} \tag{21}$$

onde $\nabla f(x(i))$ é o gradiente da função de perda, λ é o parâmetro de regularização.

Em comparação com algortimos de otimização que utilizam apenas o gradiente, levar em consideração a curvatura da superfície da função de perda é especialmente importante nos casos que a função não é convexa ou possui iterações complexas com os parâmetros de entrada. Por conseguinte, o algoritmo detém melhor performance na convergência da etapa de minimização.

3.4 **Redes Neurais**

As redes neurais são uma coleção de algoritmos modelados conforme o cérebro humano que conseguem reconhecer padrões em dados (51). As informações são interpretadas através de neurônios artificias, responsáveis por ajustar os pesos da rede. Os padrões identificados pelas redes neurais são numéricos, representados como vetores, o que permite que dados reais, como imagens, som, texto ou séries temporais sejam utilizados.

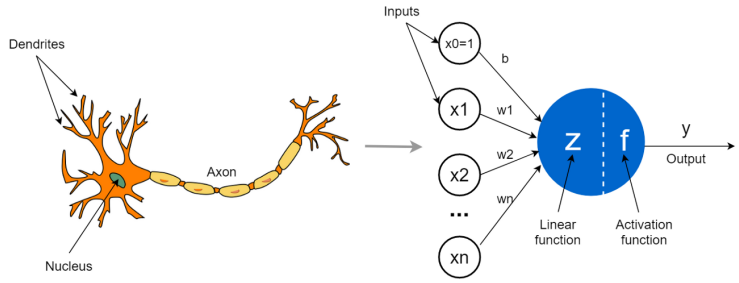
Neurônios artificias são objetos matemáticos que processam e transmitem informações através de uma rede de conexões. Cada neurônio recebe entrada de outros neurônios via sinapses e usa essa entrada para produzir uma saída. A saída é então transmitida para outros neurônios via sinapses adicionais, ver figura 14.

Os sinais de entrada em um neurônico artificial são representados por um vetor $x_i = (x_1, x_2, ..., x_n)$ e dentro do neurônico é realizado uma combinação linear com os pesos $w_i = (w_1, w_2, ..., w_n)$ da sinapse através da equação 22

$$z = \sum_{i=1}^{N} x_i w_i + b, \tag{22}$$

onde z é o potencial de ativação e b é o *bias*, um termo adicional que provê um grau de liberdade adicional. Então, uma função de ativação $f(z)$ é utilizada para decidir se o neurônio é ativado ou não. Funções de ativação comumente utilizadas são: função degrau, sigmóide, tangente hiperbólica, softmax e RELU.

Figura 14 - Comparação simplifica de um neurônio e um neurônio artificial



Legenda: A figura ilustra a evolução do conceito de neurônio tradicional para um neurônio artifical matematicamente definido.

Fonte: O autor, 2023.

3.4.1 *Perceptron*

O Perceptron é uma das implementações mais antigas e simples de neurônio artificial utilizado para aprendizado supervisionado e classificação binária. Seu objetivo é traçar uma região de decisão linear e utilizá-la para separar os dados de entrada em diferentes classes. Para isso, o algoritmo ajusta os pesos das *features* de entrada iterativamente, primeiro inicializando com pequenos valores randômicos e depois calculando os pesos conforme a equação 23

$$w_i = w_i + \eta(y - \hat{y})x_i, \tag{23}$$

onde w_i é o peso da i-ésima *feature* de entrada, η é o *learning rate*, y é a classe verdadeira do dado de entrada, $\hat{y}$ é a classe predita e x_i é o valor da i-ésima *feature* de entrada. Os pesos são recalculados até que o erro entre a classe verdadeira e a classe predita é tão pequeno quanto se queira ou um número de iterações máximo é alcançado.

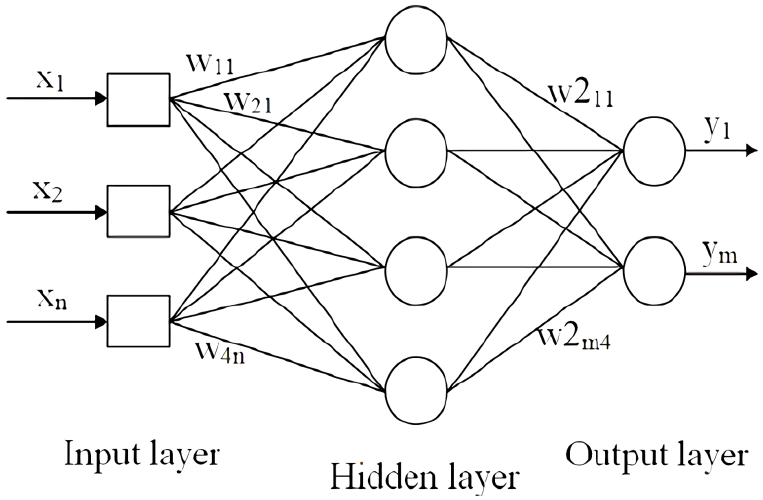
Naturalmente, como o algoritmo funciona como um classificador linear, ele diverge quando os dados não são linearmente separáveis. Nesse sentido, o algoritmo *Multi Layer Perceptron* (MLP) foi desenvolvido como a generalização em múltiplas camadas do *Perceptron* tradicional.

3.4.2 *Multi Layer Perceptron*

Multi Layer Perceptron (MLP) é um tipo de rede neural construída por múltiplas camadas de neurônios interconectados, sua topologia contém três divisões bem

definidas: A camada de entrada, uma ou mais camadas escondidas e uma camada de saída. A medida que os parâmetros de entrada atravessam as camadas, cada neurônio aplica uma transformação não linear antes de enviar o parâmetro para a próxima camada, a figura 15 ilustra a topologia de um MLP. Como a informação se move em apenas uma direção sem formar ciclos ou loops dentro da rede, esse algoritmo é classificado como *Feedforward neural network* (FNN) (52).

Figura 15 - Topologia de um Multi Layer Perceptron



Legenda: Diagrama esquemático de uma rede neural baseada no modelo *Multi Layer Perceptron*.

Fonte: ZAINAL ET AL, 2013, p. 13.

O treinamento desse algoritmo é feito através do método *backpropagation*, a ideia desse método retro propagar o erro obtido na camada de saída da rede para todas as camadas escondidas de trás para frente, atualizando os pesos de todas as camadas e consequentemente o potencial de ativação de todos os neurônios (43). O erro é calculado comparando a saída predita pela rede com a saída verdadeiro, então, o gradiente do erro é calculado tomando a derivada do erro em relação aos pesos da rede sendo usado para atualizar os pesos da rede até minimizar o erro. O processo iterativo termina quando o erro seja tão pequeno quanto se queira ou um número de iterações máximo é alcançado.

Nesse trabalho os algoritmos XGB e MLP foram estudados como propostas para criação de um discriminante entre sinal e *background* a ser aplicado sobre os dados, os parâmetros, discriminantes de cada modelo. Esses algoritmos foram escolhidos devido sua simplicidade de implementação, alta robustez e inúmeras referências de seu uso no campo aprendizado de máquina aplicado a Física de Altas Energias. Por fim, uma comparação entre os resultados dos dois modelos será apresentada no próximo capítulo.

4 ANÁLISE DE DADOS E RESULTADOS

O trabalho apresentado nesse capítulo é a contribuição do autor nos campos que concernem o *b-tagging* e aprendizado de máquina na busca por matéria escura através do processo $\bar{b}bZ(\rightarrow \ell\bar{\ell}) + \not{E}_T(a \rightarrow \chi\bar{\chi})$ (ver figura 2) utilizando todos os dados coletados durante o Run-2, isto é, com os períodos de 2016, 2017 e 2018. Em especial o período de 2016 sofreu um problema de alto depósito de energia (*High Energy Deposition Problem*, HIP) nos sensores SiStrip o que causou uma saturação temporária no módulo *front-end* APV25 e introduziu um tempo morto significativo na leitura do sistema do detector. Portanto, o período de 2016 é divido nas eras *pre* ("HIPM" ou "APV") e *post* VFP (*Preamplifier Feedback Voltage Bias*).

A probabilidade de decaimento do Higgs pesado em Za é de 43,63% e a probabilidade de decaimento do pseudoescalar a em $\chi\bar{\chi}$ é de 96,91% o que torna esse canal de decaimento excelente para o estudo do processo físico proposto. Essa análise é inédita no CERN/CMS e a determinação correta da sensibilidade do sinal pode tornar esse canal promissor no *High Luminosity* LHC. A identificação de jatos provenientes do quark bottom e a reponderação desses eventos é essencial e o uso de aprendizado de máquina se faz necessário devido à baixa seção de choque do processo de tal forma que, o uso de uma análise puramente *cut-based*[2] em contrapartida afetaria drasticamente a determinação da sensibilidade do sinal.

Em seguida, serão apresentadas as amostras de dados e amostras de simulação de Monte Carlo utilizadas no estudo, os critérios de pré-seleção dos objetos físicos e critérios de seleção de base da análise, a assinatura do sinal e a definição da região de sinal e regiões de controle dos *backgrounds* dominantes, os métodos utilizados para correção dos eventos utilizando *b-tagging* bem como o resultado das correções e por fim a implementação dos modelos de aprendizado de máquina utilizados na análise e seus discriminantes.

4.1 Amostras de dados

Todas as amostras utilizadas nesse estudo foram coletadas no formato NANOAOD (54) e processadas com um *framework* próprio desenvolvido pelo grupo que está participando desse trabalho. As amostras utilizadas para cada período estão descritas no

[2] Estratégia de análise de dados em Física Experimental de Altas Energias aonde critérios de pós-seleção são aplicados cuidadosamente com objetivo de selecionar a maior quantidade de sinal possível enquanto rejeita a maior quantidade de *background* possível.

Apêndice A.

4.2 **Amostras da simulação de Monte Carlo**

Os eventos de simulação são utilizados para modelar o processo físico estudado, dessa forma é necessário produzir os eventos de simulação específicos do canal de estudo, sinal, e também de outros processos que podem se comportar como eventos de *background*, isto é, eventos que o estado final é parecido com o sinal. As amostras de simulação utilizadas na análise referem-se à campanha de reprocessamento dos dados de 2016, 2017 e 2018, conhecido como *Ultra-Legacy*. Todas as amostras de simulação utilizadas na análise, foram produzidas pela colaboração CMS e a lista completa dos *datasets* utilizados estão presentes no apêndice B.

4.2.1 *Triggers*

É utilizado uma combinação de *triggers* de Single-Lepton (SL) e Di-Lepton (DL), de tal maneira que, *triggers* DL são responsáveis pela seleção principal e *triggers* SL ajudam a recuperar a eficiência de seleção em regiões aonde o lépton secundário tem baixo p_T e em outras topologias do evento (diferentes combinações de M_H e M_a) impondo critérios de isolamento mais relaxados. Para topologias *boosted*, as eficiências foram mensuradas utilizando o método ortogonal com respeito aos *triggers* de $\not{E}_T$ (55). Os *triggers* utilizados por período e por canal leptônico estão listados no apêndice C.

4.2.2 Critérios de pré-seleção

Alguns critérios de pré-seleção são utilizados sobre os objetos físicos e a partir disso é definido a seleção base utilizada na análise para definição da região de sinal (utilizada para o treinando dos algoritmos de aprendizado de máquina) e das regiões de controle (utilizadas para modelagem dos eventos de simulação com respeito aos dados).

4.2.2.1 Critérios de pré-seleção: Elétrons

1. $p_T > 20$ GeV
2. $|\eta| < 2{,}4$
    - Critério utilizado para seleção de objetos reconstruídos com maior precisão, é o ponto de separação físico do barril com o *endcap* do detector.
3. $|\eta| \notin [1,444; 1,566]$
4. MVA isolation ID: WP 90
    - O algoritmo MVA isolation é um algortimo multivariado que utiliza várias variáveis relacionadas ao elétron, como a energia depositada nos calorímetros e informação do sistema de trajetoria para isolar os elétrons de outras partículas. WP90 é o *Working Point* com eficiência de 90% de seleção de elétrons isolados (56).

4.2.2.2 Critérios de pré-seleção: Múons

1. $p_T > 20$ GeV
2. $|\eta| < 2{,}4$
    - Da mesma forma que os critérios de pré-seleção dos Elétrons, o critério é utilizado para seleção de objetos reconstruídos com maior precisão.
3. Muon ID: medium
    - O ID medium é otimizado para *prompt* múons e múons de decaimento de sabores pesados. Um múon "medium" é um múon relaxado com segmentos de trajetografia utilizando *hits* de mais de 80% das camadas internas de trajetografia. Se reconstruído somente como trackermuon, a compatibilidade do

segmento deve ser maior que 45,1%, caso seja reconstruído como trackermuon e globalmuon simultaneamente, a compatibilidade do segmento deve ser maior que 30,3%. O ajuste do globalmuon precisa ter $\chi^2 < 3$ e a correspondência na posição entre trackermuon e standalone muon precisa ter $\chi^2 < 12$. O χ^2 máximo calculado pelo algoritmo de detecção de vértices secundários deve ser menor que 20 (57).

4. PF isolation: tight

    - É um critério de seleção rígido aplicado aos múons baseado na informação de isolamento do algoritmo PF. O isolamento é obtido através da soma da energia depositada no cone ao redor do múon, definido pela variável ΔR ($\equiv \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2}$) e considerando o hádrons carregados e partículas neutras do PF. O WP *tight* é definido tal que a seleção de múons genuínos seja de 95% enquanto minimiza a contaminação de múons de outros jatos (57).

4.2.2.3 Critérios de pré-seleção: Jatos

1. $p_T > 20$ GeV
2. $|\eta| < 2{,}4$
3. Jet ID: TightLepVeto

    - É um algoritmo de identificação de jatos utilizado para rejeitar jatos que foram provavelmente originados do decaimento de léptons. "TightLepVeto" é o *Working Point* com objetivo de oferencer o maior nível de rejeição enquanto mantém alta a eficiência de seleção de jatos genuínos. O algoritmo utiliza o conceito de "jet-lepton overlap", isto é, a situação em que um jato é encontrado sobreposto com um elétron no evento (58) para eficientemente rejeitar prováveis jatos sobrepostos.

4. puID > tight ($p_T < 50$ GeV)

    - É um algoritmo de identificação de pileup utilizado para identificar e rejeitar jatos associados com eventos de pileup, o *Woking Point tight* oferece o maior nível de rejeição de jatos de pileup (59).

5. Jet-lepton ΔR (Isolamento mínimo do lépton com o jato mais próximo) > 0,4
6. DeepJet *b-tagging*, WP loose

4.2.2.4 Critérios de seleção base

A seleção de base é um conjunto de critérios de seleção com propósito de reduzir o *background* enquanto preserva-se a maior quantidade de sinal possível. O comportamento de algumas quantidades na seleção base, para o período de 2018, podem ser visualizadas na figura 16 e as figuras para os outros períodos estão disponíveis no apêndice D. Vale a pena ressaltar a grande sensibilidade da topologia *boosted* com $M_H = 1000$ GeV e $M_a = 100$ GeV evidente na figura 16.

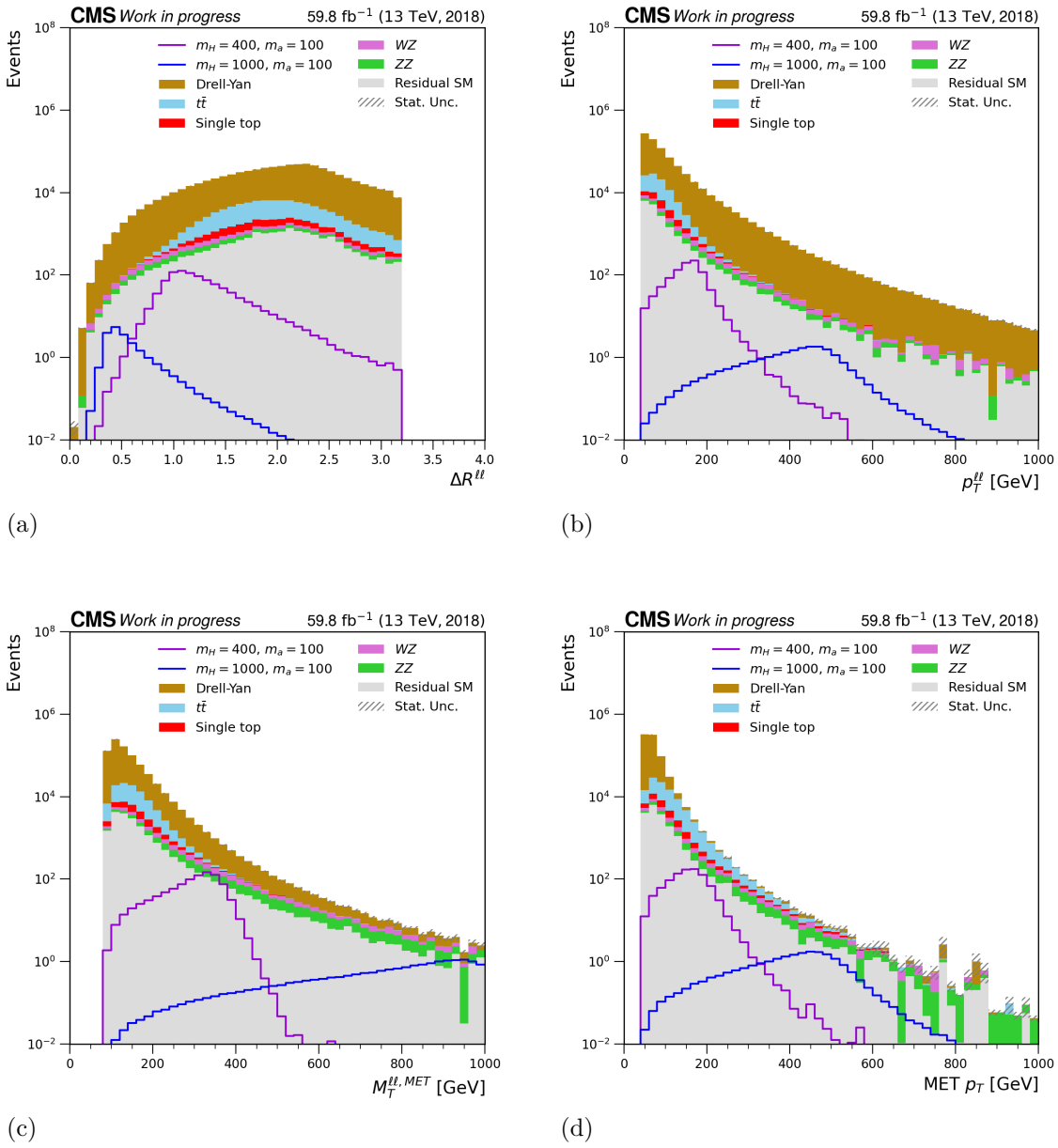
1. Léptons primários de cargas opostas
2. p_T^{ℓ} (Momentum transverso do lépton primário) > 40 GeV
3. Passar por filtros de $\not{E}_T$ (60)
4. $\not{E}_T > 50$ GeV
5. $p_T^{\ell\ell}$ (Momentum transverso do sistema de dois léptons) > 40 GeV
6. $\Delta M^{\ell\ell}$ (Diferença de massa entre sistema de dois léptons e a massa do bóson Z) $>$ 25 GeV
7. $\Delta R^{\ell\ell}$ (Cone de isolamento do sistema de dois léptons) $< 3{,}2$
8. $\Delta\phi^{\ell\ell,\not{E}_T}$ (Diferença do ângulo polar entre sistema de dois léptons e $\not{E}_T$) $> 0{,}8$
9. $M_T^{\ell\ell,\not{E}_T}$ (Massa transversa do sistema de dois léptons e $\not{E}_T$) > 90 GeV

4.2.3 Sinal

A assinatura do sinal utilizada para definir a região de sinal é dada pelo estado final com a presença de pares de léptons de cargas opostas, energia trasnversa faltante, presença de pelo menos um jato proveniente do quark bottom. Apesar do diagrama de Feynman, ver figura 2, do processo requerer a presença de dois jatos provenientes do quark bottom, nesse estudo optamos por um critério de seleção mais relaxado, isto é, a presença de pelo menos um jato proveniente do quark bottom. Essa escolha foi motivada pela preservação do número de eventos de sinal disponíveis para um melhor treinamento dos algoritmos de aprendizado de máquina.

Os pontos de sinal foram definidos para diferentes combinações de M_H e M_a. Os efeitos cinemáticos devido à variação dos parâmetros do modelo ($\tan\beta$, $\sin\theta$, y_x e etc.) foram armazenados como pesos nas amostras de Monte Carlo. Os pontos de sinal

Figura 16 - ΔR do sistema de dois léptons, p_T do sistema de dois léptons, massa transversa do sistema de dois léptons e $\not{E}_T$ e $\not{E}_T$ na seleção base para o período de 2018



(a) (b)

(c) (d)

Legenda: As figuras acima apresentam as distribuições de (a) $\Delta R^{\ell\ell}$, (b) $p_T^{\ell\ell}$, (c) $M_T^{\ell\ell,\not{E}_T}$ e (d) $\not{E}_T$. na seleção base.

Fonte: O autor, 2023.

propostos, ver tabela 1, para sondagem foram escolhidos de forma a maximizar o espaço de parâmetros a ser explorado. Grandes valores para diferença de massa entre M_H e M_a classificam o ponto de sinal como *boosted* (ou topologia *boosted*), isto é, grande parte da diferença de massa é convertida em forma de momentum para os léptons produzidos no decaimento do bóson Z através da conservação de energia e momentum, ver figura 17.

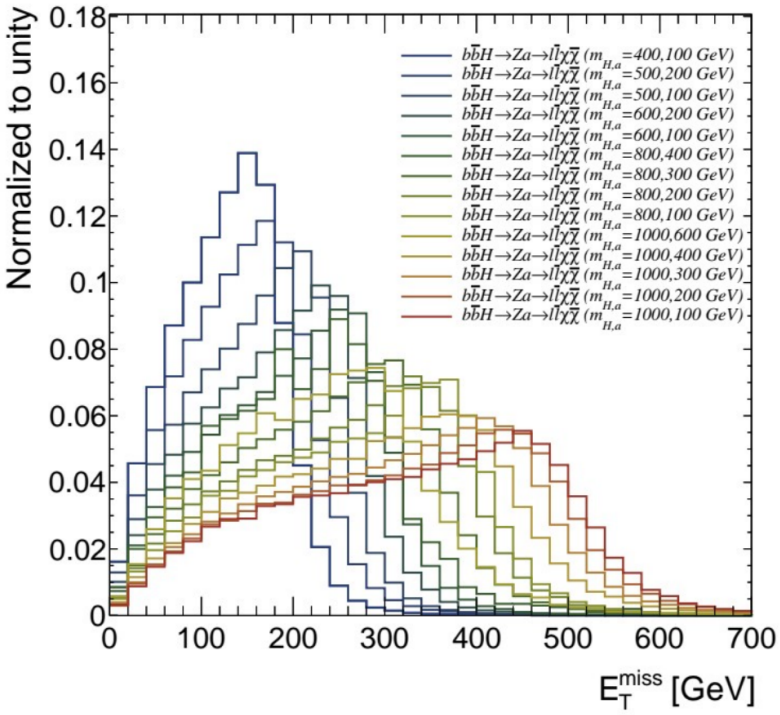
Tabela 1 - Pontos de sinal propostos para sondagem

$M_H[GeV]/M_a[GeV]$	100	200	300	400	600	800
400	✓	✓	×	×	×	×
500	✓	✓	✓	×	×	×
600	✓	✓	✓	✓	×	×
800	✓	✓	✓	✓	✓	×
1000	✓	✓	✓	✓	✓	✓

Legenda: A primeira coluna da tabela são os possíveis valores de M_H e a primeira linha são os possíveis valores de M_a, cada célula ij da tabela é um par $M_H - M_a$ proposto para sondagem. Se o valor da célula é um ✓ o ponto de sinal será sondado e caso seja um × não será sondado.

Fonte: O autor, 2023.

Figura 17 - Comparação da energia transversa perdida dos pontos de sinal propostos



Legenda: Energia transversa perdida a nível de gerador dos pares de pontos de sinal propostos para sondagem. A medida que a diferença de massa entre H e a aumentam (topologias *boosted*) a energia transversa perdida aumenta linearmente.

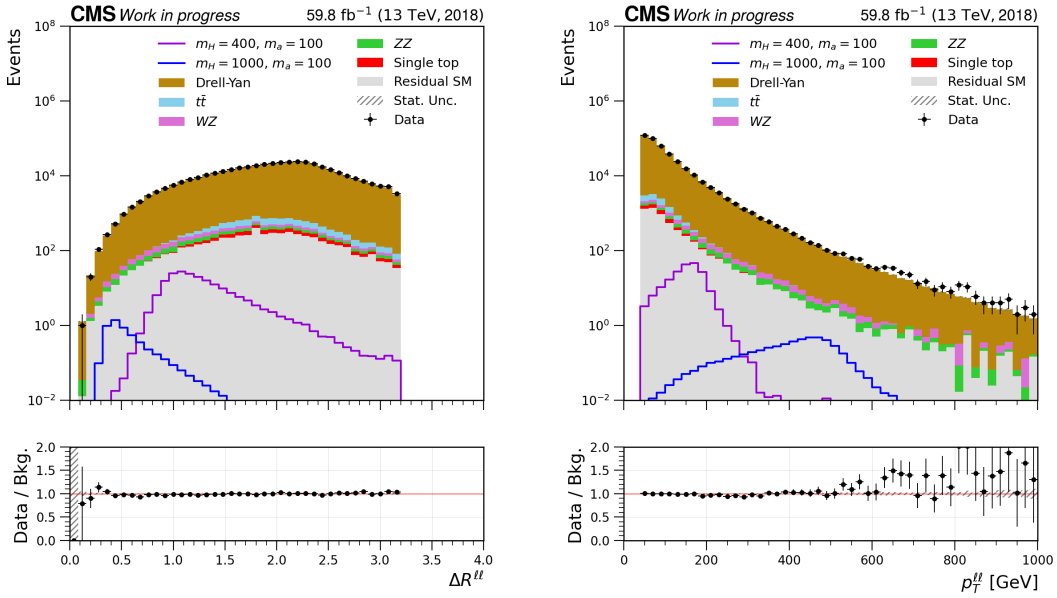
Fonte: PÉREZ ADÁN, 2021.

4.2.4 *Drell-Yan* (DY)

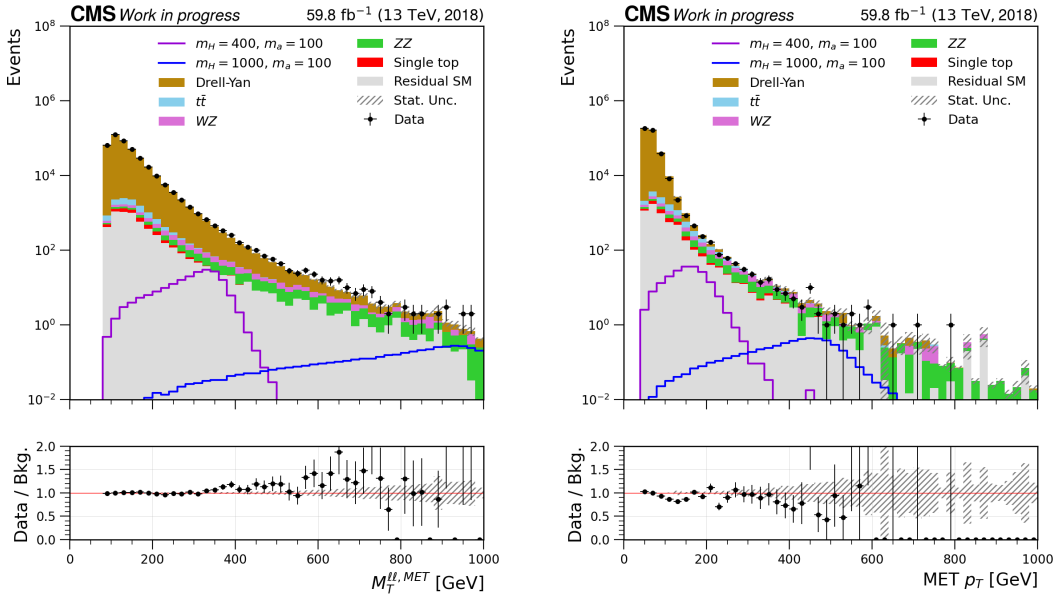
O processo de Drell-Yan consiste na produção de pares de léptons através da aniquilação quark-antiquark no decaimento de um fóton virtual Z/γ^*. É uma das maiores fontes de *background* da análise devido à alta seção de choque e a presença de pares de léptons de cargas opostas no estado final. Para verificar a boa modelagem dos eventos de simulação foi definida uma região de controle com pureza de 97% para o período de 2018 que pode ser visualiza na figura 18 (as figuras para os outros períodos estão disponíveis no apêndice E), em cima dos critérios de seleção base, com a seguinte configuração:

- $N_{bjets} = 0$
- Presença de um par de léptons com mesmo sabor e cargas opostas.

Figura 18 - ΔR do sistema de dois léptons, p_T do sistema de dois léptons, massa transversa do sistema de dois léptons e $\not{E}_T$ e $\not{E}_T$ na região de controle do DY para o período de 2018



(a) (b)



(c) (d)

Legenda: As figuras acima apresentam as distribuições de (a) $\Delta R^{\ell\ell}$, (b) $p_T^{\ell\ell}$, (c) $M_T^{\ell\ell,\not{E}_T}$ e (d) $\not{E}_T$. na região de controle do DY.

Fonte: O autor, 2023.

4.2.5 $t\bar{t}$

A produção de pares de léptons proveniente de pares de quark-top é outro *background* de maior contribuição da análise, além da produção leptônica esse processo de decaimento também produz pares de jatos provenientes do quark bottom e energia transversa perdida proveniente dos neutrinos. Assim como processo de Drell-Yan possui alta seção de choque, contudo, seu estado final é ainda mais similar com o sinal. A definição da região de controle com pureza de 87% para o período de 2018 que pode ser visualizada na figura 19 (as figuras para os outros períodos estão disponíveis no apêndice F) e considera os seguintes critérios:

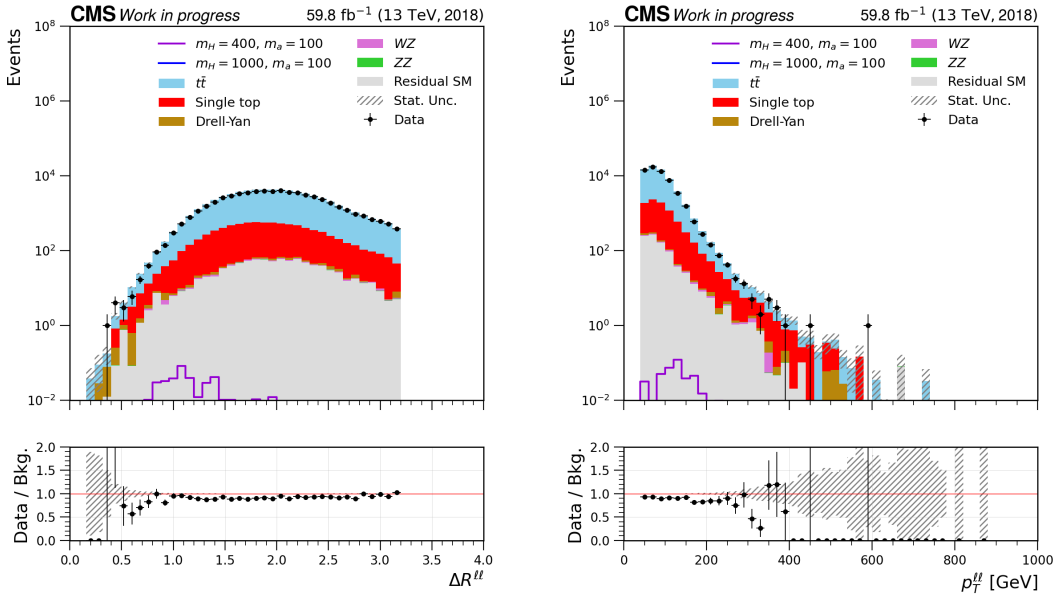
- $N_{bjets} \geq 1$
- Presença de um par de léptons com diferentes sabores e cargas opostas.

4.2.6 WZ

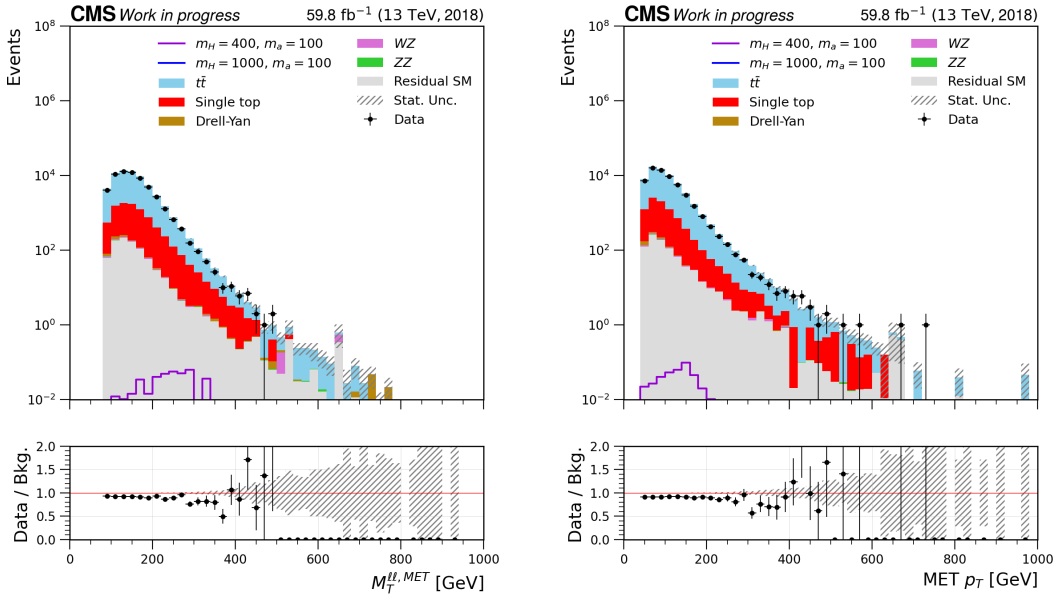
Um processo de decaimento menos dominante na análise é dos bósons WZ, enquanto o bóson Z produz um par de léptons de cargas opostas, o bóson W produz um lépton e um neutrino. A configuração do estado final é similar ao sinal, com a presença de pelo menos dois léptons de cargas opostas e energia transversa perdida. Definimos a região de controle com pureza de 55% para o período de 2018 que pode ser visualizada na figura 20 (as figuras para os outros períodos estão disponíveis no apêndice G) com os seguintes critérios:

- $N_{bjets} = 0$
- Presença de um par de léptons (ℓ_1 e ℓ_2) com mesmo sabores e cargas opostas.
- Presença de um lépton extra (ℓ_3)
- $60 < M_T^{\not{E}_T,\ell_3} < 100$ GeV

Figura 19 - ΔR do sistema de dois léptons, p_T do sistema de dois léptons, massa transversa do sistema de dois léptons e $\not{E}_T$ e $\not{E}_T$ na região de controle do $t\bar{t}$ para o período de 2018



(a) (b)

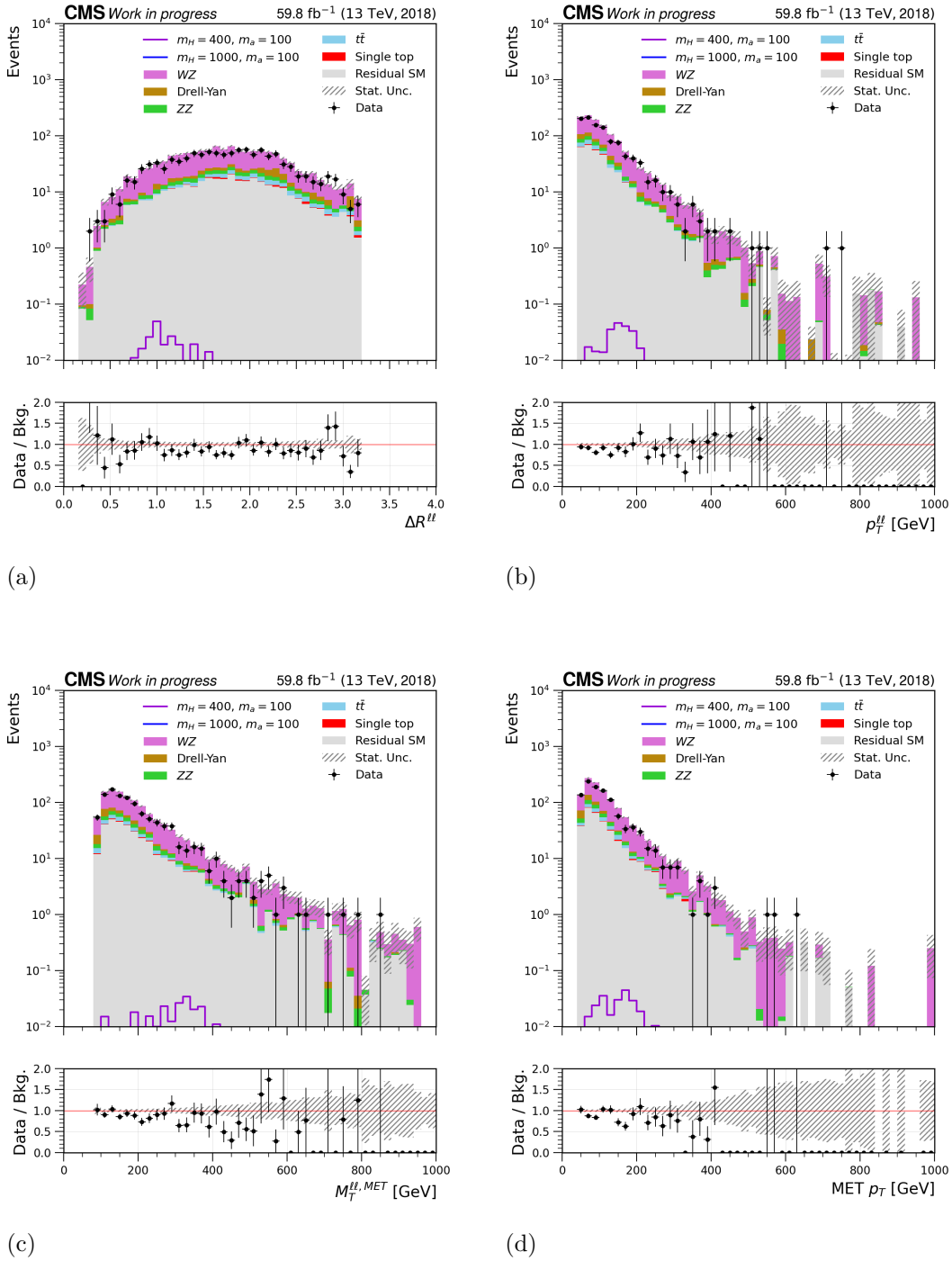


(c) (d)

Legenda: As figuras acima apresentam as distribuições de (a) $\Delta R^{\ell\ell}$, (b) $p_T^{\ell\ell}$, (c) $M_T^{\ell\ell,\not{E}_T}$ e (d) $\not{E}_T$. na região de controle do $t\bar{t}$.

Fonte: O autor, 2023.

Figura 20 - ΔR do sistema de dois léptons, p_T do sistema de dois léptons, massa transversa do sistema de dois léptons e $\not{E}_T$ e $\not{E}_T$ na região de controle do WZ para o período de 2018



(a) (b)

(c) (d)

Legenda: As figuras acima apresentam as distribuições de (a) $\Delta R^{\ell\ell}$, (b) $p_T^{\ell\ell}$, (c) $M_T^{\ell\ell,\not{E}_T}$ e (d) $\not{E}_T$. na região de controle do WZ.

Fonte: O autor, 2023.

4.2.7 ZZ

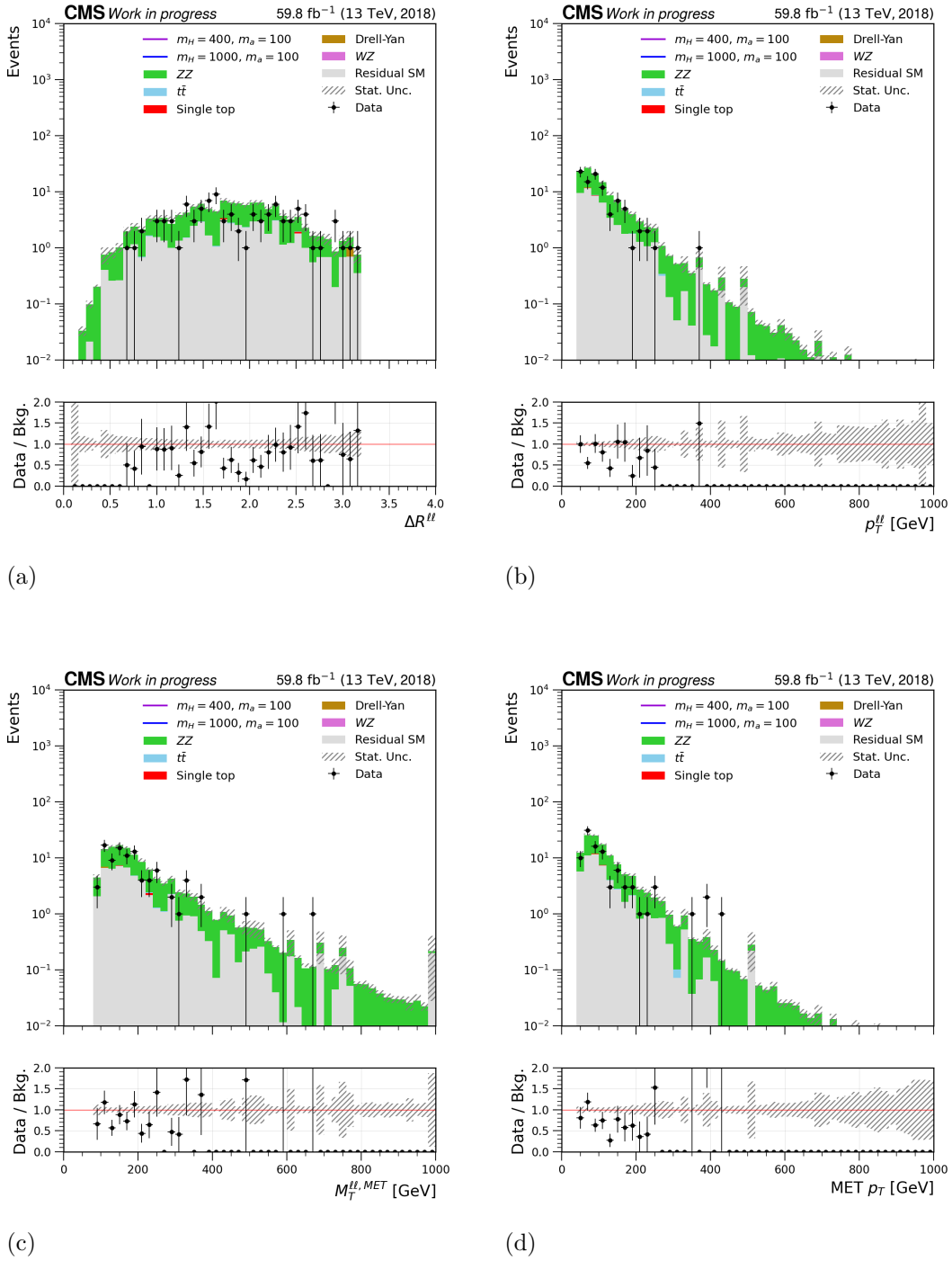
O processo de decaimento de pares de bósons Z produz quatro léptons no estado final e em algumas situações os eventos podem ser mal reconstruídos devido à complexidade de identificação dos dois pares de léptons e o processo contribui como um *background* menos dominante na análise. A região de controle com pureza de 58% para o período de 2018 que pode ser visualizada na figura 21 (as figuras para os outros períodos estão disponíveis no apêndice H) foi definida com os seguintes critérios de seleção:

- $N_{bjets} = 0$
- Presença de um par de léptons (ℓ_1 e ℓ_2) com mesmo sabor e cargas opostas.
- Presença de um par extra de léptons (ℓ_3 e ℓ_4) com mesmo sabor e cargas opostas.
- $M^{\ell_3\ell_4} > 10$ GeV

4.2.8 Processos residuais

Muitos outros processos de decaimento produzem *backgrounds*, pouco semelhantes com o sinal e possuem baixa seção de choque, como outras composições de bósons vetoriais (WZZ, WWZ, WWW e WW), outros processos envolvendo o quark-top junto de bósons vetoriais (TWZ, TTWZ, TTWW, TTZZ) e outros processos. Devido à baixa contribuição estatística desses conjuntos de dados, esses processos residuais previstos no Modelo Padrão foram agrupados no grupo “Residual SM” não foi definida uma região de controle para o grupo.

Figura 21 - ΔR do sistema de dois léptons, p_T do sistema de dois léptons, massa transversa do sistema de dois léptons e $\not{E}_T$ e $\not{E}_T$ na região de controle do ZZ para o período de 2018



(a) (b)

(c) (d)

Legenda: As figuras acima apresentam as distribuições de (a) $\Delta R^{\ell\ell}$, (b) $p_T^{\ell\ell}$, (c) $M_T^{\ell\ell,\not{E}_T}$ e (d) $\not{E}_T$. na região de controle do ZZ.

Fonte: O autor, 2023.

4.3 **Correção dos eventos utilizando b-tagging**

Para que a simulação modele bem os dados, garantido que o valor da seção de choque utilizado esteja correto, ela precisa ser corrigida para apresentar uma razão entre o número de eventos de dados e eventos de simulação o mais próxima possível da unidade. A correção dos eventos é a combinação de correções de diversos objetos físicos, como, *Jet Energy Scale* (JES) (62), *Jet Energy Resolution* (JER) (63), *b-tagging* (64), MET Corrections (65), Muon Rochester (66), Prefiring (67), Jet Pileup ID (68) e outros. Todas essas correções são utilizadas pela colaboração CMS sendo recomendado que sejam implementadas na análise de cada grupo, contudo, nessa seção será apresentado apenas os métodos que o autor utilizou para corrigir as eficiências dos algoritmos de *b-tagging* como contribuição para a análise de dados realizada pelo grupo.

4.3.1 Escolha do algoritmo de *b-tagging* e *Working Point*

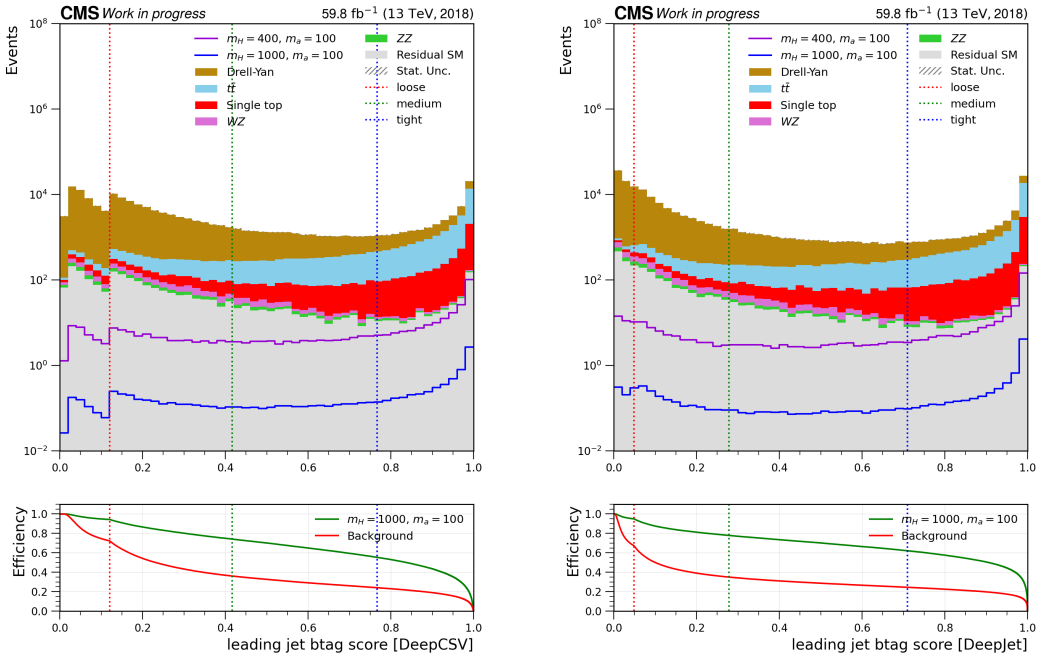
Atualmente o grupo de *b-tagging* do CMS recomenda o uso dos *taggers* DeepCSV ou DeepJet, a escolha do algoritmo e *Working Point* depende do que melhor se encaixa na análise, portanto, para essa análise foi escolhido a combinação que maximiza a eficiência de seleção de sinal e *background*. O processo de decaimento em estudo possui baixa seção de choque, portanto, para o desenvolvimento de um algoritmo de aprendizado de máquina eficiente faz-se necessário preservar a maior quantidade de eventos de sinal possível sem prejudicar o balanço de eventos de dados e simulação. O gráfico de eficiência (b) na figura 22 mostra que a maior eficiência de seleção de eventos de sinal é utilizando o *tagger* DeepJet com WP *loose* para o período de 2018. A comparação entre os dois algoritmos para os demais períodos estão disponíveis no apêndice I.

4.3.2 Reponderação dos eventos

O grupo de *b-tagging* do CMS mensura a eficiência dos fatores de escala (SF) de *b-tagging* para jatos provenientes de quark bottom (b), quark charm (c) jatos leves (udsg). A reponderação do peso dos eventos considera esses fatores de escala e a eficiência do *tagger* nas amostras de simulação, que deve ser calculada pelo analista, pois dependem da cinemática do evento (p_T, η, sabor do jato) (69).

Para essa análise foi utilizado o método de reponderação utilizando os fatores de escala e eficiências de *b-tagging* no MC recomendado pelo grupo de *b-tagging* do CMS, o objetivo dessa abordagem é prever com precisão os eventos nos dados, ajustando apenas os pesos dos eventos de simulação selecionados, sem precisar adicionar eventos que não

Figura 22 - Comparação dos algoritmos DeepCSV e DeepJet na região de sinal para o período de 2018



(a) (b)

Legenda: Comparação do valor do discriminante dos algoritmos (a) DeepCSV e (b) DeepJet para o jato de quark b com maior valor do discriminante nos *Working Points* loose, medium e tight. O gráfico de eficiência embaixo de cada discriminante representa a eficiência de seleção de sinal/*background* se um corte fosse aplicado em um determinado valor do discriminante.

Fonte: O autor, 2023.

passaram na seleção. Isso elimina problemas com variáveis indefinidas em etapas posteriores da análise, como a massa ou comprimento de decaimento de um vértice secundário se ele não for reconstruído (69). Além disso, vale ressaltar que este método não resulta em migração de eventos entre bins de multiplicidade de jatos provenientes do quark bottom. O cálculo da correção considerando o uso de um WP é:

$$P(MC) = \sum_{i=tagged} \epsilon_i \sum_{j=nottagged} (1 - \epsilon_j) \tag{24}$$

$$P(DATA) = \sum_{i=tagged} SF_i\epsilon_i \sum_{j=nottagged} (1 - SF_j\epsilon_j) \tag{25}$$

$$w = \frac{P(DATA)}{P(MC)} \tag{26}$$

onde ϵ_i é a eficiência de *b-tagging* nas amostras de MC para jatos provenientes do quark bottom, ϵ_j é a eficiência de *b-tagging* nas amostras de MC para jatos não provenientes do quark bottom, SF_i a eficiência dos fatores de escala e w o peso do evento corrigido. Portanto, faz-se necessário calcular os mapas de eficiência de *b-tagging* das amostras para o cálculo da correção.

4.3.3 Mapas de eficiência

Os mapas de eficiência são calculados em função da cinemática de cada processo de decaimento e produzem um valor de eficiência no espaço de fase (p_T, η) para cada sabor de jato (ver figura 23). Os bins de η são construídos em espaços iguais {0-0,6; 0,6-1,2; 1,2-2,4}, enquanto a determinação correta dos bins de p_T afetam diretamente a suavidade das curvas de eficiência. Dessa forma, um algoritmo foi desenvolvido[3] para calcular a distribuição ótima de bins de p_T com base na incerteza do número de eventos dentro do intervalo de bin analisado, isto é, um bin só é aceito se a incerteza dos eventos contidos no bin é menor que um valor fixo. A eficiência bin a bin é calculada de acordo com

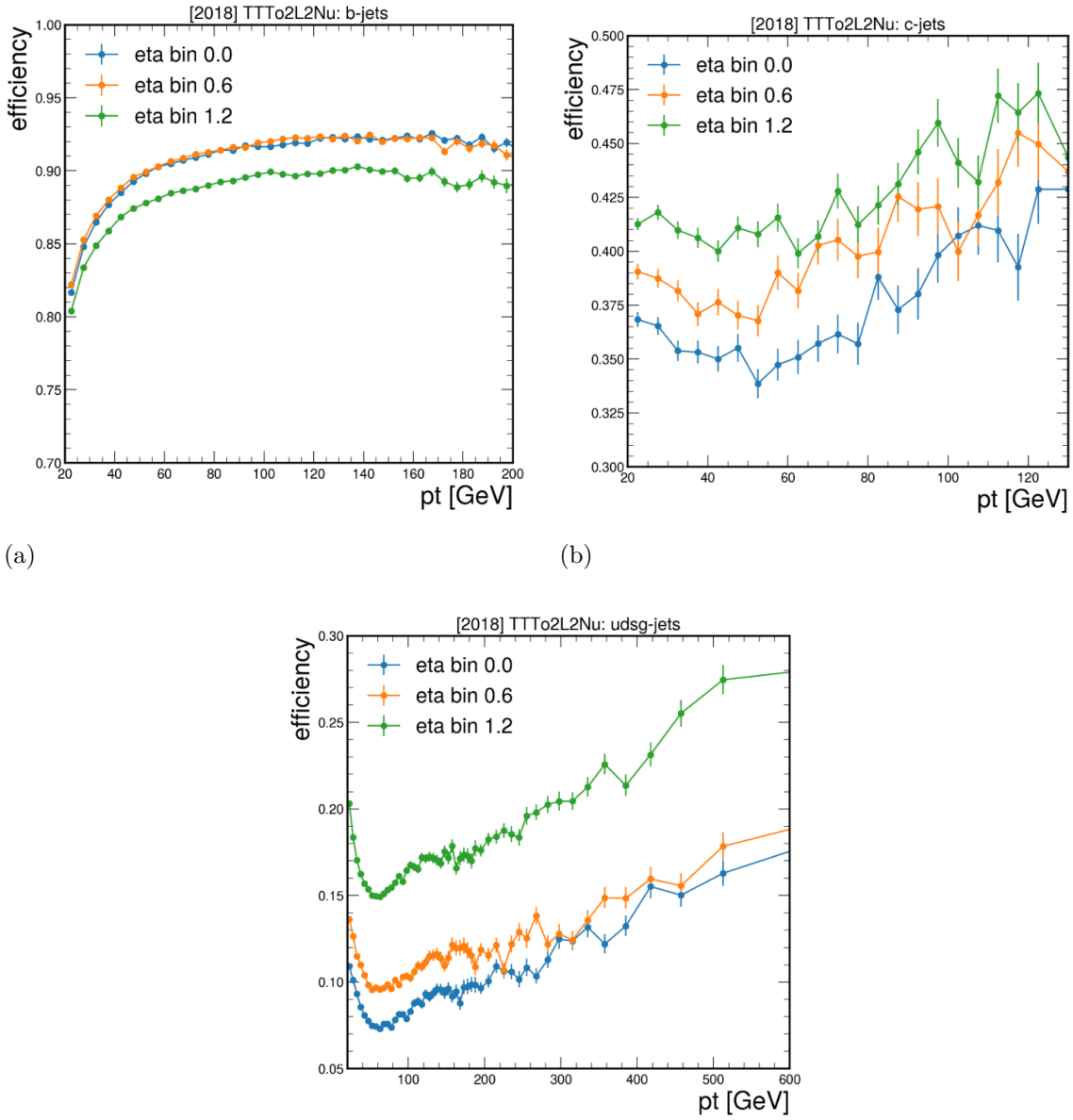
$$\epsilon_f(i,j) = \frac{N_f^{b-tagged}(i,j)}{N_f^{total}(i,j)} \tag{27}$$

aonde $N_f^{b-tagged}(i,j)$ e $N_f^{total}(i,j)$ são, respectivamente, o número total e o número de jatos *b-tagged* com sabor f no espaço (p_T, η) no bin (i, j). O sabor do jato é determinado utilizando o valor verdadeiro (*truth*[4]) do gerador de Monte Carlo (70).

[3] btaggingEffMaps (https://github.com/gabrielmscampos/btaggingEffMaps) para produção dos mapas de eficiência e btageffanalyzer (https://github.com/gabrielmscampos/btageffanalyzer) para inserção dos mapas de eficiência no *framework* de análise do grupo em C++

[4] O valor original de um observável é referido como *truth* e o valor medido pelo detector é referido como reconstruído. Na geração de amostras de Monte Carlo, o gerador produz valores reconstruídos e guarda um mapa desses valores para o valor *truth*.

Figura 23 - Mapas de eficiência para a amostra de simulação do processo $t\bar{t} \to \ell\bar{\ell} + \not{E}_T(\nu)$



(a) (b)

(c)

Legenda: Mapas de eficiência para a amostra de simulação do processo $t\bar{t} \to \ell\bar{\ell} + \not{E}_T(\nu)$ com bins de p_T escolhidos dinamicamente pelo algoritmo de otimização, onde (a) é o eficiência do *tagger* para b-jets, (b) é a eficiência do *tagger* para c-jets e (c) é a eficiência do *tagger* para udsg-jets.

Fonte: O autor, 2023.

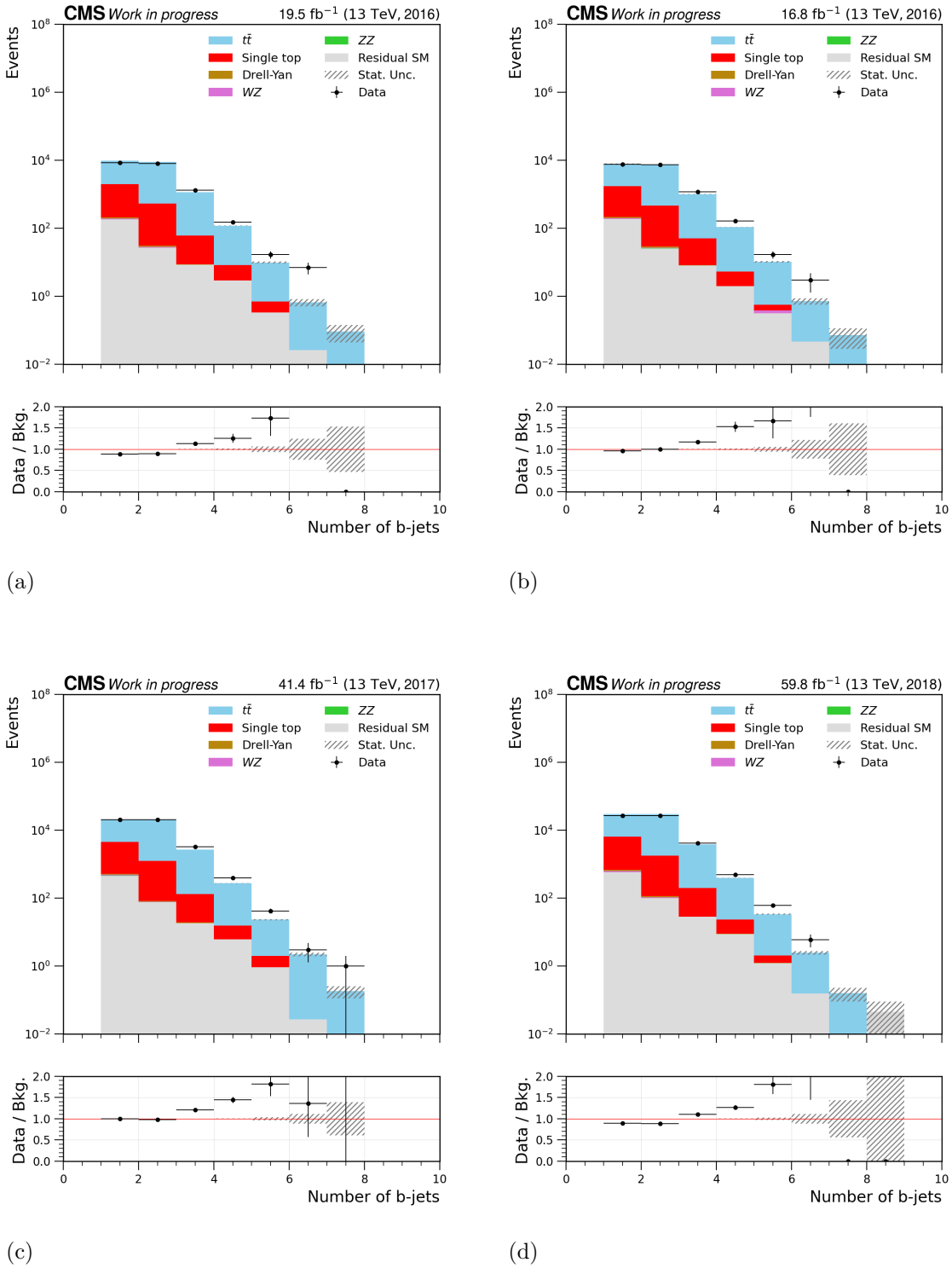
4.3.4 Correção dos eventos

Com uso dos mapas de eficiência construídos para essa análise e a equação 26 os eventos são corrigidos. A figura 25 mostra a comparação entre os dados e eventos de

simulação na multiplicidade de jatos provenientes do quark bottom. É possível observar como o balanço é drasticamente melhorado com a atualização dos pesos comparando a multiplicidade de jatos provenientes de quark bottom antes (figura 24) e depois (figura 25) das correções.

Uma anomalia nas correções foi encontrada para o período 2016 post-VFP devido a um erro no cálculo da eficiência dos fatores de escala dos jatos leves fornecidos pelo grupo de *b-tagging* do CMS para esse período no WP *loose* e afeta ambos os algoritmos de identificação de jatos de quark b, DeepJet e DeepCSV. Foi recomendado pelo grupo de *b-tagging* do CMS que fosse usado os fatores de escala do período 2016 pre-VFP enquanto uma correção não fosse fornecida.

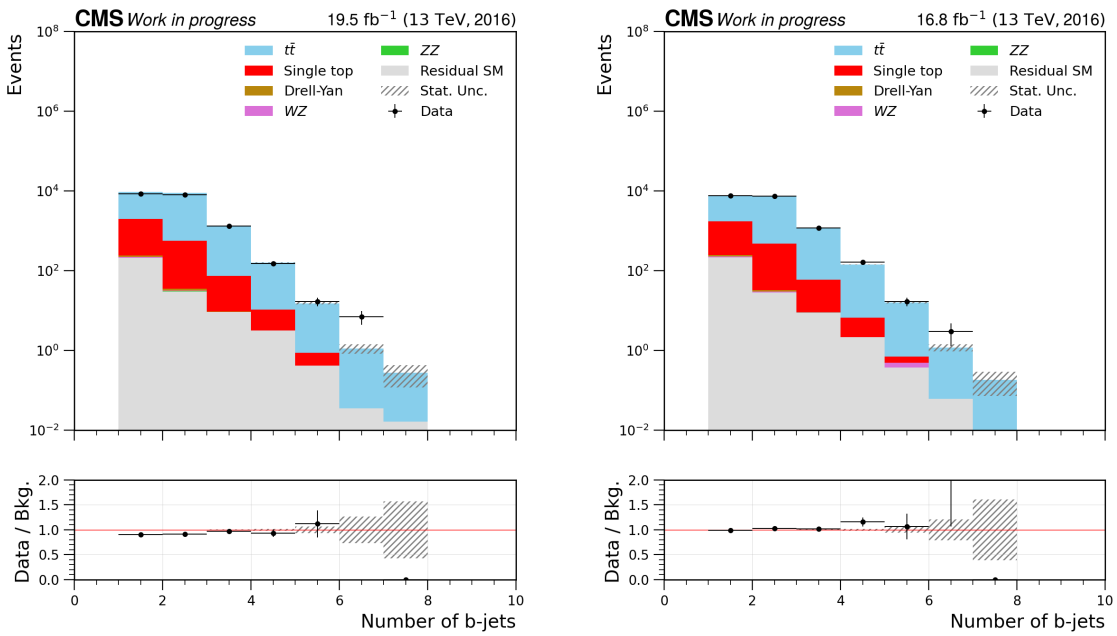
Figura 24 - Multiplicidade de jatos provenientes do quark bottom antes da correção do b-tagging



Legenda: Distribuições da multiplicade de jatos provenientes do quark bottom utilizando o algoritmo DeepJet antes de corrigir os eventos com métodos de *b-tagging*, onde (a) é a distribuição para o periodo de 2016 pre-VFP, (b) a distribuição para 2016 post-VFP, (c) a distribuição para 2017 e (d) para distribuição de 2018.

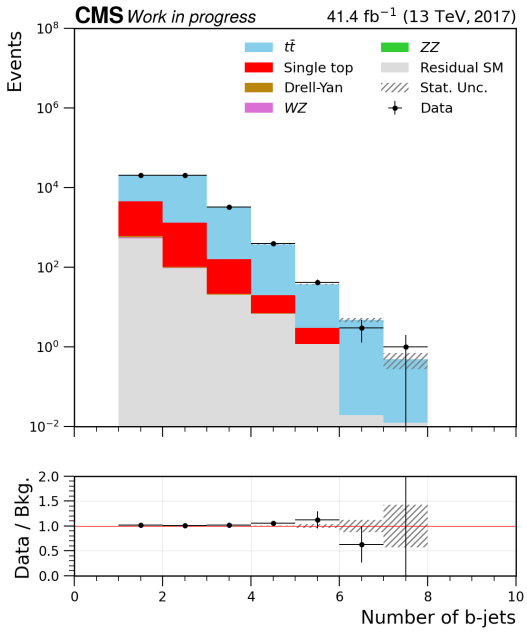
Fonte: O autor, 2023.

Figura 25 - Multiplicidade de jatos provenientes do quark bottom depois das correções do b-tagging

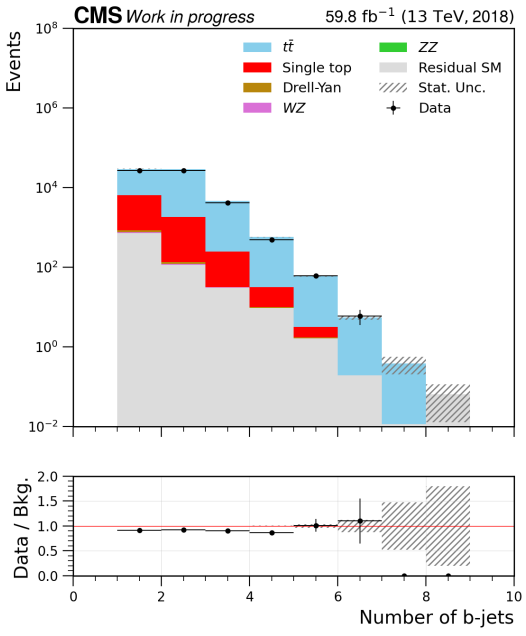


(a)

(b)



(c)



(d)

Legenda: Distribuições da multiplicade de jatos provenientes do quark bottom utilizando o algoritmo DeepJet depois de corrigir os eventos com métodos de *b-tagging*, onde (a) é a distribuição para o periodo de 2016 pre-VFP, (b) a distribuição para 2016 post-VFP, (c) a distribuição para 2017 e (d) para distribuição de 2018.

Fonte: O autor, 2023.

4.4 O uso do Aprendizado de Máquina na seleção de eventos de sinal

A proposta para esse trabalho é utilizar aprendizado de máquina e produzir uma variável discriminadora entre o sinal e o *background* sem impor nenhuma restrição física no processo de descoberta de padrões. O discriminante produzido será utilizado em um momento posterior para o cálculo do limite de exclusão do sinal. Para isso, as *features* são escolhidas com base no seu poder discriminativo e o balanço entre os dados e eventos de simulação. Os conjuntos de dados de sinal e *background* são cuidadosamente divididos em *datasets* de treino e teste de modo que o peso do conjunto de eventos de simulação agrupado (DY, $t\bar{t}$, ZZ, WZ, Redisual) é normalizado, de acordo com

$$norm\_factor(train) = \frac{\sum_i w_i^{full}}{\sum_i w_i^{train}} \tag{28}$$

$$w_i^{train} = \frac{w_i^{train}}{norm\_factor(train)} \tag{29}$$

onde, $norm\_factor(train)$ é o fator de normalização do conjunto de treino definido em função da soma dos pesos de todo o conjunto de dados (w^{full}) e a soma dos pesos do conjunto de treino (w^{train}) e w_i^{train} é o i-ésimo peso de um evento. Esse procedimento é realizado antes da etapa de treinamento para que a seção de choque de cada processo físico seja respeitada.

Além disso, um estudo foi realizado por um membro do grupo dessa análise para verificar a possibilidade de construir um discriminante agrupando todos os pontos de sinal de tal forma que simplificaria o cálculo das regiões de decisão (71). A conclusão desse estudo foi que nenhuma eficiência no discriminador era perdida para topologias *boosted* (subseção 4.2.3) e outras topologias aonde a diferença de massa entre M_H e M_a era pequena, a perda de eficiência era desprezível. Portanto, os modelos apresentados nesse trabalho foram treinados agrupando cuidadosamente todos os pontos de sinal de tal maneira que, o peso de cada ponto sinal fosse normalizado (de forma similar a 29) antes do agrupamento para que todos os pontos de sinal tivessem a mesma importância para o modelo.

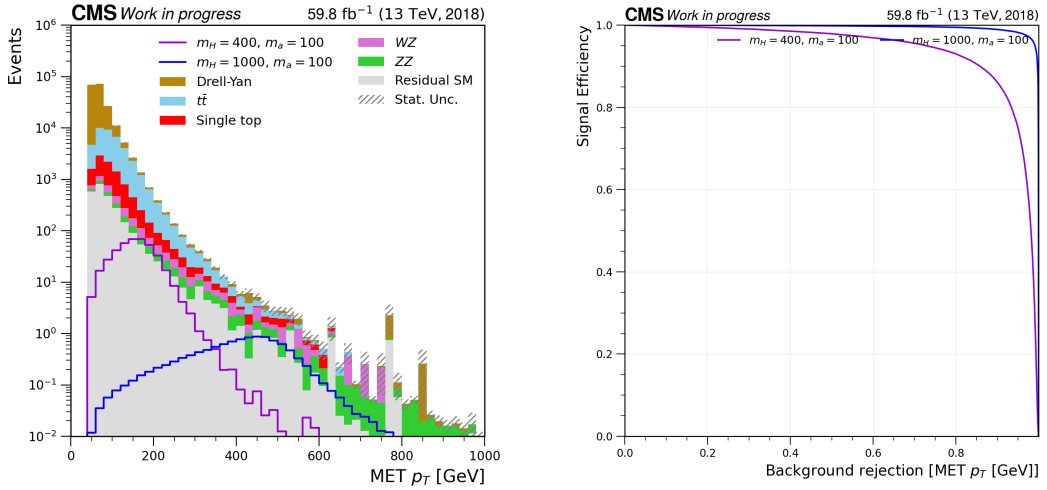
4.4.1 Seleção de *features*

As *features* são variáveis que descrevem um objeto e carregam informação relevante para etapa de aprendizado e são selecionadas conforme o poder discriminatório

da variável. As distribuições de energia transversa perdida (figura 26) e momentum transverso do sistema de dois léptons (figura 27) na região de sinal possuem excelente poder discriminativo como pode ser visto analisando a curva ROC das respectivas figuras e sinais de topologia *boosted* são identificados facilmente devido o grande momentum presente no estado final com relação ao *background*. Na figura (a) da figura 27 é possível observar a fácil separabilidade entre sinal e *backgound* no ponto de sinal *boosted* $M_H = 1000$ GeV e $M_a = 100$ GeV.

Por outro lado, as grandezas hadrônicas não oferecem poder discriminativo comparado às leptônicas (figura 28), contudo, carregam informações sobre a topologia do evento e expõe a característica frontal dos jatos de sinal (figura 29), isto é, jatos com altos valores de $|\eta|$ que são detectados em regiões mais afastadas da região central do detector. Por esse motivo, o número de jatos provenientes do quark bottom é uma *feature* selecionada. A lista de *features* utilizadas está disponível no apêndice K e o comportamento das *features* para os outros períodos estão disponíveis no apêndice J.

Figura 26 - Energia transversa perdida na região de sinal para o período de 2018



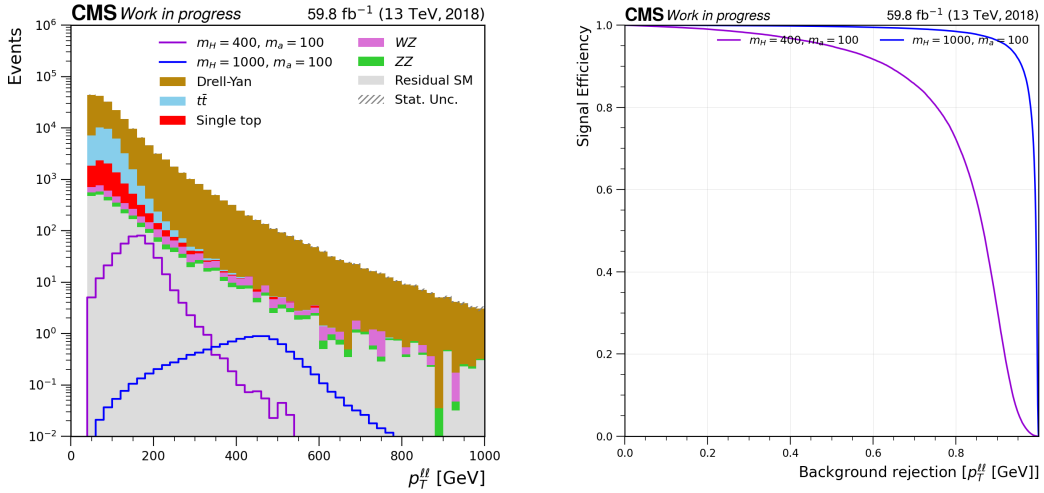
(a) (b)

Legenda: (a) Distribuição de energia trasversa perdida e (b) curva ROC. Quanto maior a eficiência na seleção de sinal e rejeição de *background* na cuva ROC, maior é o poder discriminativo.

Fonte: O autor, 2023.

Além disso, a correlação entre as *features* é avaliada através de uma matriz de correlação, ver figura 30, para garantir que múltiplas variáveis com a mesma informação física não sejam usadas simultâneamente na etapa de treinamento.

Figura 27 - Momentum transverso do sistema de dois léptons na região de sinal para o período de 2018

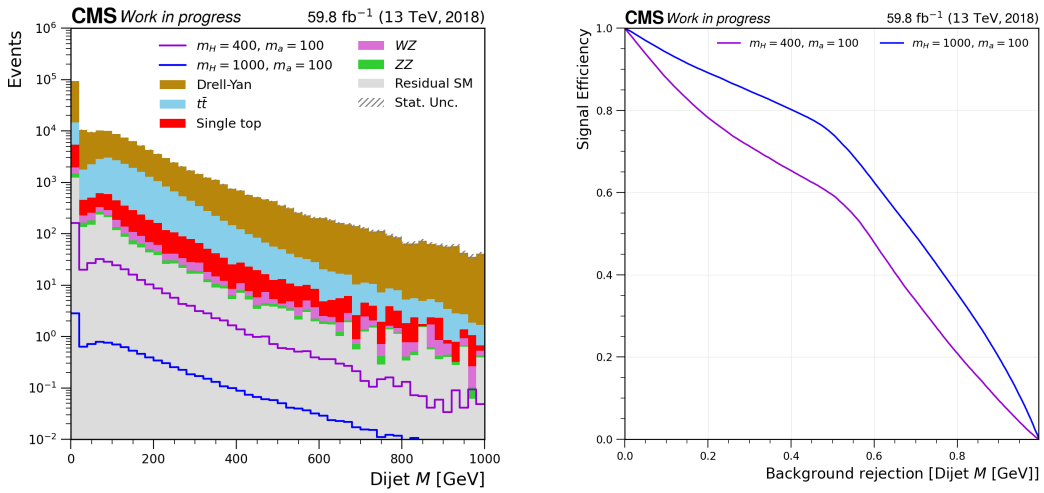


(a) (b)

Legenda: (a) Distribuição do momentum transverso do sistema de dois léptons e (b) curva ROC. Assim como na distribuição de energia transversa perdida, a separabilidade dentre sinal e *background* para topologia *boosted* $M_H = 1000$ GeV e $M_a = 100$ GeV é evidente.

Fonte: O autor, 2023.

Figura 28 - Massa reconstruída do sistema de dois jatos na região de sinal para o período de 2018

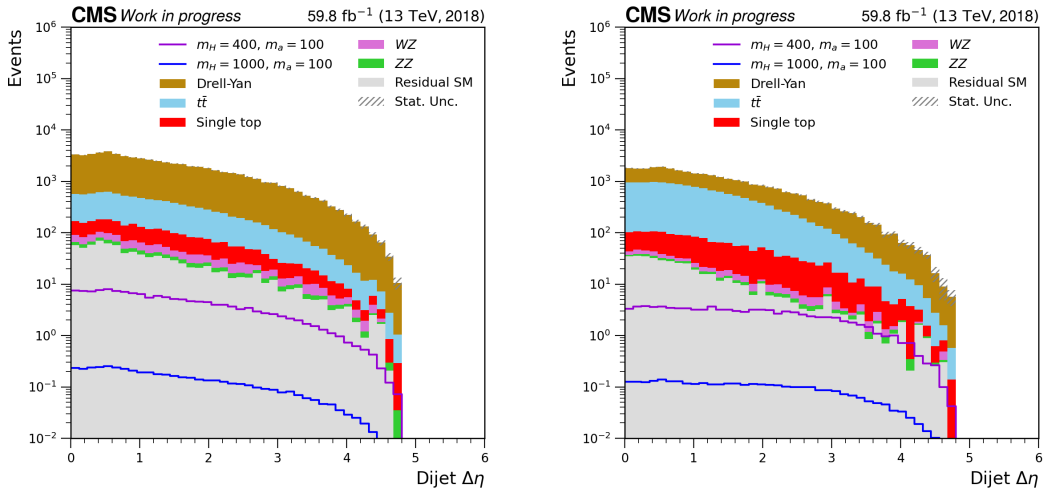


(a) (b)

Legenda: (a) Distribuição da massa reconstruída do sistema de dois jatos (b) curva ROC. A figura (b) demonstra o pouco poder discriminativo obtido de grandezas hadrônicas devido a posição central da curva ROC.

Fonte: O autor, 2023.

Figura 29 - Distribuição do $\Delta\eta$ do sistema de dois jatos na região de sinal para o período de 2018

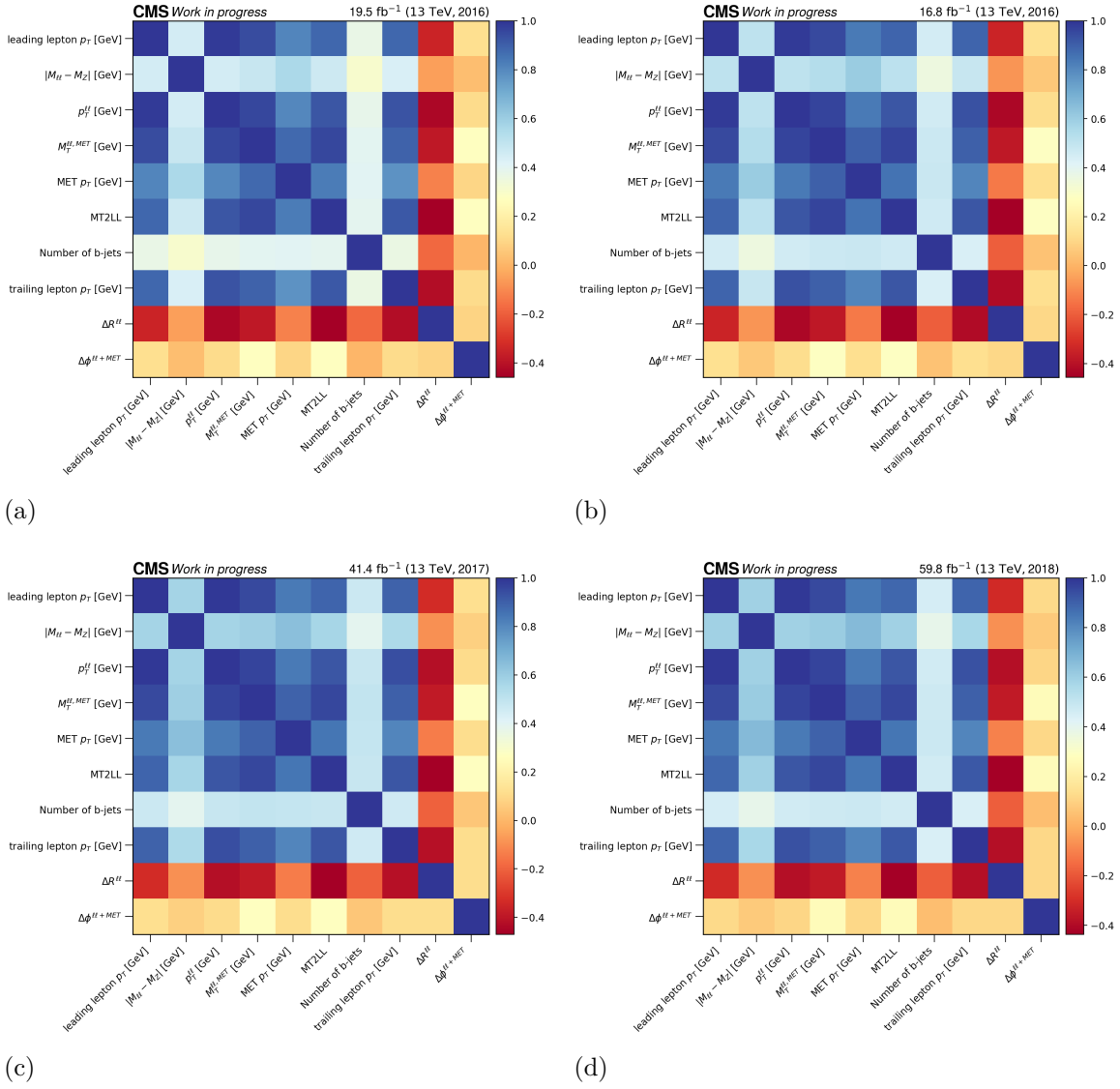


(a) (b)

Legenda: (a) $\Delta\eta$ entre 1 b-jet e qualquer outro jato e (b) $\Delta\eta$ entre 2 b-jets. O comportamento frontal pode ser visto na figura (b), aonde as distribuições de sinal se estendem quase que paralelamente ao eixo X do gráfico até altos valores de $\Delta\eta$, enquanto a distribuições do *background* se curva.

Fonte: O autor, 2023.

Figura 30 - Matriz de correlação das *features* utilizadas



(a) (b)

(c) (d)

Legenda: Matrizes de correlação para cada período, onde (a) 2016 pre-VFP, (b) 2016 post-VFP, (c) 2017 e (d) 2018.

Fonte: O autor, 2023.

4.4.2 XGBoost (XGB)

O XGBoost (XGB) foi implementado nesse trabalho utilizando a biblioteca em Python com o mesmo nome. O modelo foi escolhido devido a sua velocidade e performance no treinamento de grandes conjuntos de dados, além de oferecer excelente robustez. Os melhores hiperparâmetros são escolhidos por uma busca randômica em uma rede de hiperparâmetros utilizando o método *k-fold cross validation*. A ideia por trás deste método é dividir os dados em k "pedaços", onde cada pedaço é usado como conjunto de teste uma vez e como conjunto de treinamento k-1 vezes. O modelo é treinado em k-1 das dobras e, em seguida, testado no pedaço restante. Este processo é repetido k vezes, com cada pedaço sendo usado como conjunto de teste uma vez. A métrica de desempenho final é a média das métricas de desempenho calculadas em cada iteração. Este método ajuda a mitigar o problema de *overfitting*, pois usa todos os dados tanto para treinamento quanto para teste. Ele também pode dar uma estimativa mais robusta do desempenho do modelo, pois executa vários testes em diferentes subconjuntos de dados.

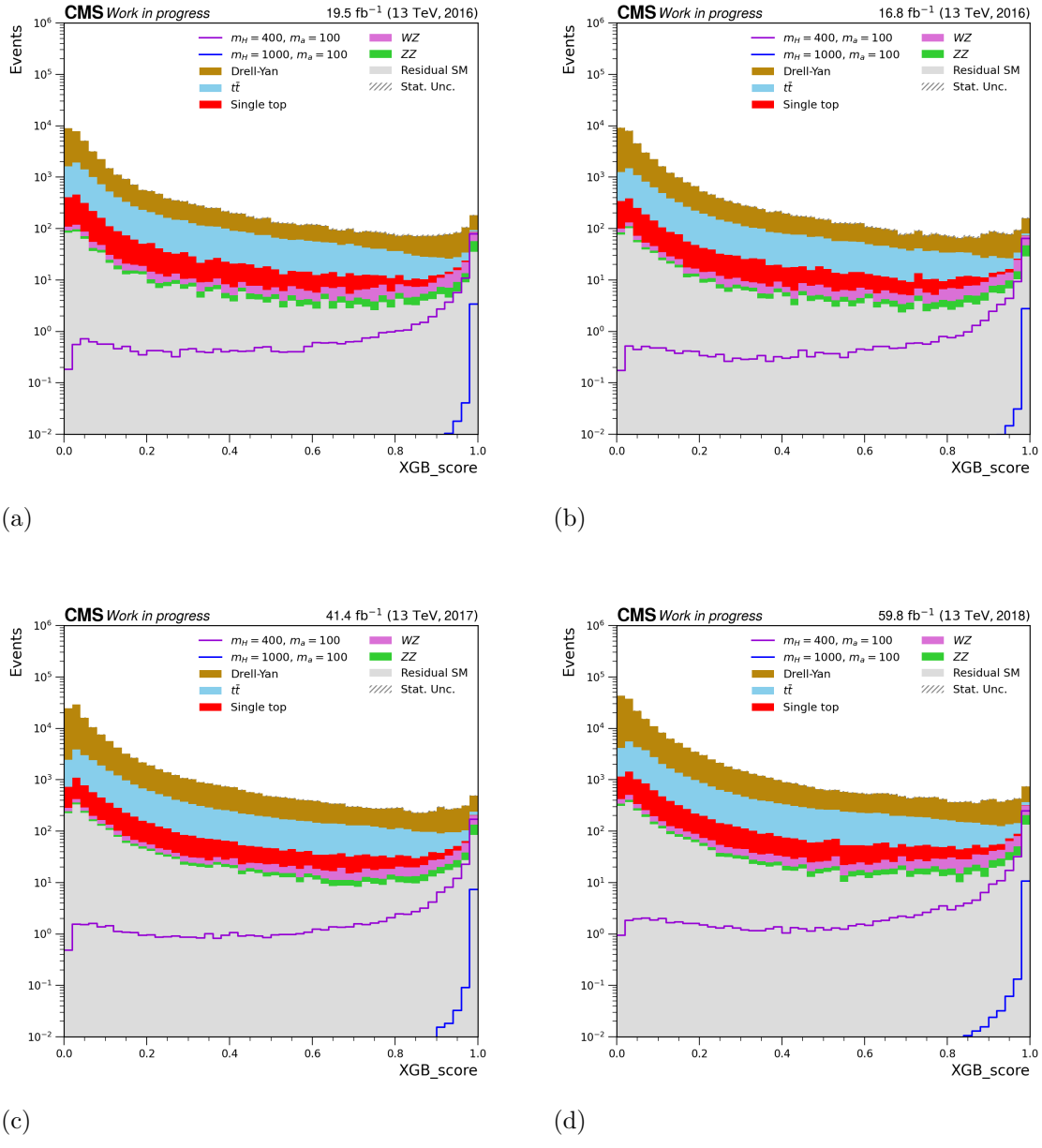
Além do *k-fold cross validation* para mitigar o *overfitting*, a correspondência entre os discriminantes nos *datasets* de treino e teste também são avaliados. A tabela 2 relaciona os hiperparâmetros utilizados para treinamento do modelo e a figura 31 apresenta o discriminante do melhor modelo, onde é possível observar que o pico da distribuição de sinal é bem distinto em relação ao *background*.

Tabela 2 - Hiperparâmetros utilizados no modelo XGB para todos os períodos

Hiperparâmetro	Descrição	Valor
train_size	Tamanho do *dataset* de treino	0,7
n_estimators	Número de árvores decisão que compõe o *ensemble*	500
min_child_weight	Soma mínima do peso da instância necessária em um nó filho	4
learning_rate	Tamanho do passo usado na atualização para evitar o overfitting	0,01
subsample	Proporção de subamostra das instâncias de treinamento	0,60
colsample_bytree	É a proporção de subamostra de colunas ao construir cada árvore	0,80
max_depth	Profundidade máxima de uma árvore	5
eval_metric	Métrica de avaliação para o conjunto de validação	error

Fonte: O autor, 2023.

Figura 31 - Discriminantes dos modelos baseados em XGB para todos os períodos



(a) (b)

(c) (d)

Legenda: Discriminantes produzidos utilizando o modelo XGB para os períodos (a) 2016 pre-VFP, (b) 2016 post-VFP, (c) 2017 e (d) 2018.

Fonte: O autor, 2023.

4.4.3 *Multi Layer Perceptron* (MLP)

Outro modelo, baseado em *Multi Layer Perceptron*, foi proposto para comparar com os resultados obtidos pelo XGBoost. O MLP foi implementado utilizando a biblioteca Keras (72) e Tensorflow (73) e o *overfitting* foi controlado utilizando o método *early-stopping* monitorando o valor da *loss function* a cada época em um *batch* do conjunto de dados de teste. A ideia por trás deste método é interromper o processo de treinamento quando o desempenho em um conjunto de validação deixa de melhorar.

Diferente do XGB a busca pelos melhores hiperparâmetros e topologia da rede é por força bruta. É definida uma rede de hiperparâmetros e topologias e para cada combinação é treinado um modelo diferente, foi necessário em torno de 3000 processos para determinação dos melhores hiperparâmetros. Então, o melhor conjunto de hiperparâmetros (ou melhor modelo) é escolhido maximizando o produto da eficiência e pureza da performance do modelo, pois em um problema de seleção de eventos buscamos minimizar o erro estatístico na determinação da seção de choque (74), através de

$$\frac{1}{\Delta {\sigma_s}^2} = \frac{1}{\sigma_s^2}\epsilon_s \rho S_{tot}, \tag{30}$$

onde σ_s é a seção de choque, ϵ_s é a eficiência, ρ é a pureza e S_{tot} o número total de eventos de sinal. O conjunto de dados de treino e teste foi normalizado utilizando o Z-Score[5] ponderado pelo peso normalizado dos eventos. Esse procedimento garante que o formato das distribuições físicas permaneça o mesmo (preservando a característica do objeto físico) e é essencial no treinamento de redes neurais do tipo MLP (75).

A tabela 3 descreve a topologia e hiperparâmetros da rede utilizados e a tabela 4 relaciona o ponto de parada do algoritmo *early stopping* para cada período. A figura 32 apresenta a evolução da *loss function* durante o treinamento e o critério de parada utilizado no algoritmo de *early stopping* foi quando o treinamento atingisse o menor valor da *loss function*. A figura 33 o discriminante produzido pelo modelo MLP, que assim como XGB, tem uma excelente performance na discriminação de sinal e fundo.

[5] É uma maneira de descrever um ponto de dados em termos de sua relação com a média e o desvio padrão de um grupo de pontos, $z_i = \frac{x_i - \bar{x}}{\sigma}$.

Tabela 3 - Hiperparâmetros utilizados e topologia da rede do modelo MLP para todos os períodos

Hiperparâmetro	Descrição	Valor
n_layers	Número de camadas escondidas	2
units	Número de neurônios em cada camada escondida	20
hidden_activation	Função de ativação em cada camada escondida	elu
output_acitvation	Função de ativação na camada de saída	sigmoid
optmizer	Algoritmo utilizado para minimização da *loss function*	Adam
optmizer_lr	Tamanho do passo usado na atualização para evitar o overfitting	0,01
loss	*Loss function*	Binary Crossentropy (BCE)
batch_size	Número de amostras por atualização de gradiente	1000
epochs	Número de épocas para treinar o modelo	2000
patiente	Número de épocas sem melhora *loss function* no qual o treinamento será interrompido	300
train_size	Tamanho do *dataset* de treino	0,5

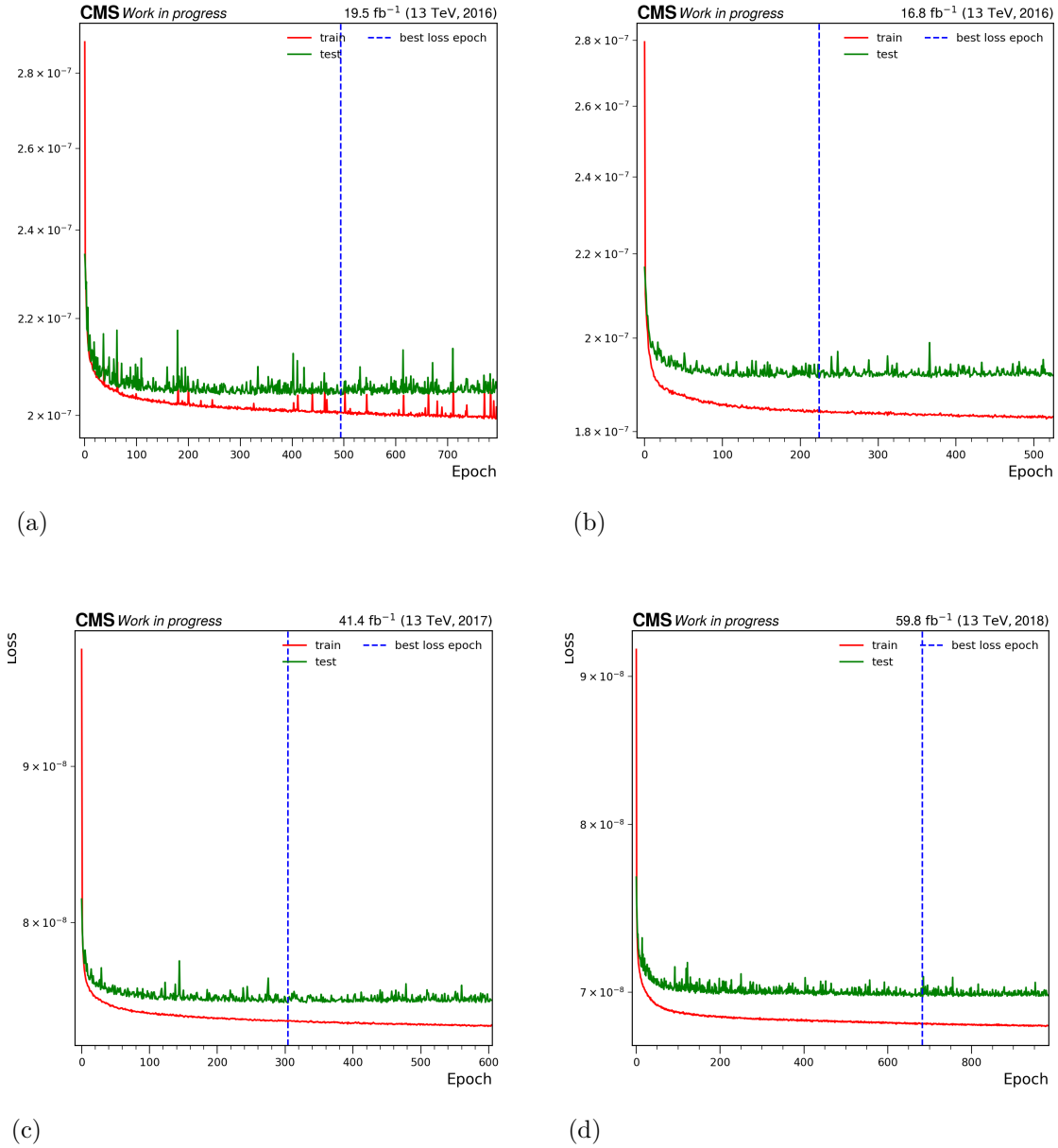
Fonte: O autor, 2023.

Tabela 4 - Ponto de parada do algoritmo de *early-stopping*

Período	Época	*Loss* no treino	*Loss* no teste	Acurácia no treino	Acurácia no teste
2016 pre-VFP	494	$2{,}01 \times 10^{-7}$	$2{,}04 \times 10^{-7}$	0,82	0,82
2016 post-VFP	224	$1{,}84 \times 10^{-7}$	$1{,}91 \times 10^{-7}$	0,82	0,82
2017	304	$7{,}42 \times 10^{-8}$	$7{,}53 \times 10^{-8}$	0,84	0,84
2018	683	$6{,}83 \times 10^{-8}$	$6{,}97 \times 10^{-8}$	0,83	0,82

Fonte: O autor, 2023.

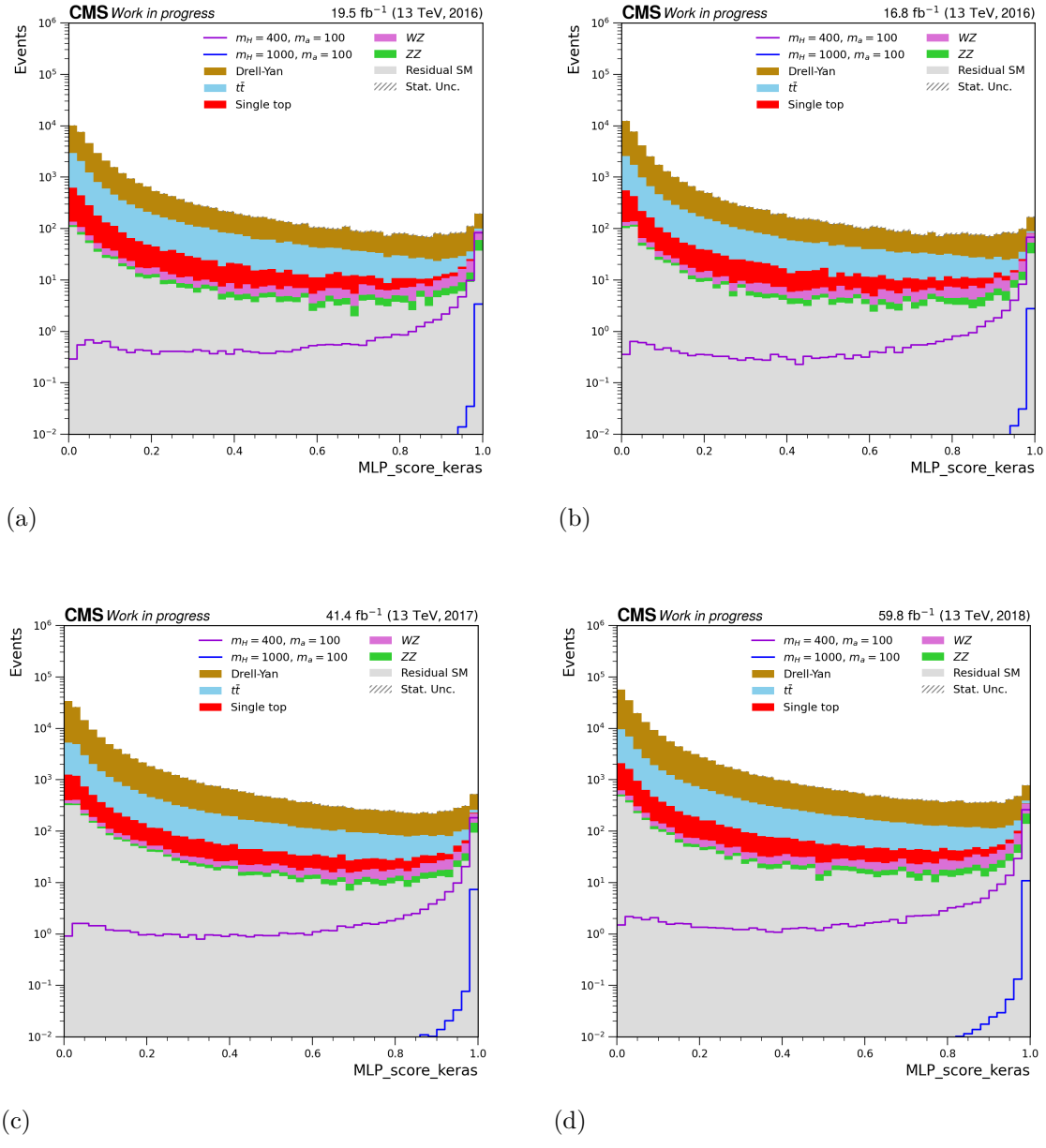
Figura 32 - *Loss function* de todos os períodos durante o treinamento do modelo MLP



Legenda: Evolução da *loss functions* no treinamento do modelo MLP para os períodos (a) 2016 pre-VFP, (b) 2016 post-VFP, (c) 2017 e (d) 2018. A linha tracejada em cada figura representa a época onde a *loss function* apresentou o menor valor.

Fonte: O autor, 2023.

Figura 33 - Discriminantes dos modelos baseados em MLP para todos os períodos



Legenda: Discriminantes produzidos utilizando o modelo MLP para os períodos (a) 2016 pre-VFP, (b) 2016 post-VFP, (c) 2017 e (d) 2018.

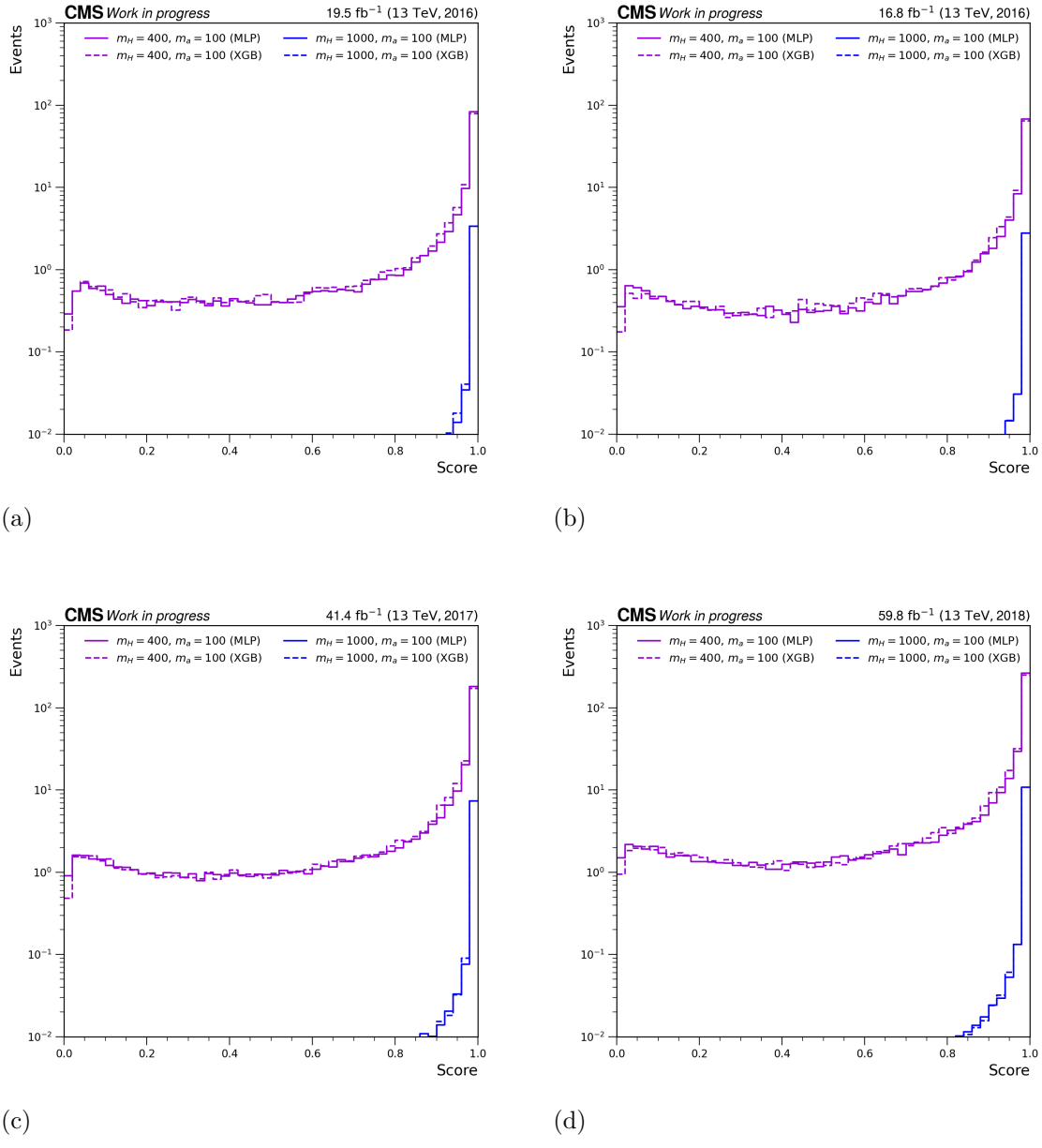
Fonte: O autor, 2023.

4.4.4 Comparação entre os modelos propostos

Uma comparação entre os modelos XGB e MLP para os pontos de sinal $M_H, M_a = (400, 100)$ GeV e $M_H, M_a = (1000, 100)$ GeV é apresentada na figura 34, aonde é possível concluir que os dois modelos possuem essencialmente a mesma performance. Ainda que o XGB tenha uma pequena vantagem na topologia *boosted* essa vantagem é desprezível quando consideramos o número de eventos disponíveis.

O modelo MLP foi implementado na análise de dados devida sua maior capacidade de melhoria através do uso de outras técnicas de redes neurais profundas, como adição de camadas convolucionais e recorrentes. O estudo entre o XGB e o MLP foi essencial na tomada de decisão de qual modelo utilizar e atuou como uma validação cruzada de métodos de diferentes de aprendizado de máquina, que, demonstrando comportamento similar, trouxe confiança na implementação do método escolhido na análise. O código para produção dos resultados apresentados e treinamento dos modelos propostos está disponível em (76).

Figura 34 - Comparação entre os discriminantes para todos os períodos



(a) (b)

(c) (d)

Legenda: Comparação entre os discriminantes utilizando os modelos XGB e MLP para os períodos (a) 2016 pre-VFP, (b) 2016 post-VFP, (c) 2017 e (d) 2018.

Fonte: O autor, 2023.

CONCLUSÃO

O trabalho aqui apresentado foi realizado no estudo da busca por matéria escura fermiônica através do processo de decaimento além do Modelo Padrão do bóson de Higgs pesado do modelo 2HDM produzindo o estado final $\bar{b}bZ(\rightarrow \ell\bar{\ell}) + \not{E}_T(a \rightarrow \chi\bar{\chi})$ utilizando dados do CMS/LHC, implementando métodos para aplicação de fatores de correção para jatos provenientes de quark bottom e modelos de aprendizado de máquina para produção de um discriminante entre sinal e *background*.

A correção dos eventos foi imprescindível para continuidade da análise, devido grande impacto da reponderação dos eventos com métodos de *b-tagging* e a necessidade de identificação de jatos provenientes do quark bottom para definição da região de sinal e regiões de controle. Os algoritmos de classificação propostos, XGBoost e Multi Layer Perceptron, demonstraram comportamento similar, com uma pequena vantagem (desprezível devido à estatística disponível) para o XGB na discriminação de pontos de sinai *boosted* ainda assim, a performance para topologias *boosted* de sinal dos dois modelos é excelente devido a evidente separabilidade observada nas grandezas leptônicas e de energia transversa perdida.

O trabalho aqui apresentado foi inserido na análise de dados em andamento, que está seguindo os passos para uma possível publicação. Portanto, pode-se concluir que o objetivo desse trabalho foi alcançado.

Outro trabalho desenvolvido pelo autor durante o Mestrado foi a contribuição no grupo AlCaDB do CMS durante os preparativos para tomada de dados de 2022 (Run-3), detalhes sobre esse trabalho estão descritos no Apêndice L.

REFERÊNCIAS

1 GONçALVES, D.; MACHADO, P. A.; NO, J. M. *Simplified models for dark matter face their consistent completions. Physical Review D*, [S.l.], American Physical Society (APS), v. 95, n. 5, 2017. ISSN 2470-0029.

2 TUNNEY, P.; NO, J. M.; FAIRBAIRN, M. *Probing the pseudoscalar portal to dark matter via* $\bar{b}bZ(\rightarrow \ell\ell)+ \not{E}_T$: From the LHC to the Galactic Center excess. *Phys. Rev. D*, [S.l.], v. 96, n. 9, p. 095020, 2017.

3 SEIDEN, A. *Particle physics*: a comprehensive introduction. [S.l.]: Addison Wesley, 2005.

4 PERKINS, D. *Introduction to High Energy Physics*. [S.l.]: Cambridge University Press, 2000. ISBN 9780521621960.

5 ANDERNACH, H.; ZWICKY, F. *English and Spanish Translation of Zwicky's (1933) The Redshift of Extragalactic Nebulae*. [S.l.]: arXiv, 2017.

6 BOSMA, A. *The Distribution and kinematics of neutral hydrogen in spiral galaxies of various morphological types*. 1978. Tese (Doutorado) — Universidade de Groningen, Holanda, 1978.

7 RUBIN, V. C.; FORD W. KENT, J. *Rotation of the Andromeda Nebula from a Spectroscopic Survey of Emission Regions. Astrophysical Journal*, [S.l.], v. 159, p. 379, 1970.

8 GREEN, A. *Dark matter in astrophysics/cosmology. SciPost Physics Lecture Notes*, [S.l.], Stichting SciPost, 2022.

9 SANDERS, R. H.; MCGAUGH, S. S. *Modified Newtonian Dynamics as an Alternative to Dark Matter. Annual Review of Astronomy and Astrophysics*, [S.l.], Annual Reviews, v. 40, n. 1, p. 263–317, 2002.

10 STEFANIAK, T. *Higgs physics beyond the Standard Model*. [S.l.]: arXiv, 2019.

11 SERVANT, G.; TULIN, S. *Baryogenesis and Dark Matter through a Higgs Asymmetry. Physical Review Letters*, [S.l.], American Physical Society (APS), v. 111, n. 15, 2013.

12 CERDENO, D. G. *WIMPs*: A brief bestiary. *In: PATRAS WORKSHOP ON AXIONS, WIMPS AND WISPS (AXION-WIMP 2008), 4.*, 2008. Hamburg, DE. Proceedings ... Hamburg, 2008. P. 9–12.

13 ARCADI, G.; DJOUADI, A.; RAIDAL, M. *Dark Matter through the Higgs portal. Physics Reports*, [S.l.], Elsevier BV, v. 842, p. 1–180, 2020.

14 PECCEI, R. D. *QCD, Strong CP and Axions*. [S.l.: s.n.]. 1996.

15 JAECKEL, J.; RYBKA, G.; WINSLOW, L. *Axion Dark Matter*. [S.l.]: arXiv, 2022.

16 GUNION, J. F.; HABER, H. E. *CP-conserving two-Higgs-doublet model*: The approach to the decoupling limit. *Physical Review D*, [S.l.], American Physical Society (APS), v. 67, n. 7, 2003. ISSN 1089-4918.

17 JENNICHES, L.; STURM, C.; UCCIRATI, S. *Electroweak corrections in the 2HDM for neutral scalar Higgs-boson production through gluon fusion. Journal of High Energy Physics*, [S.l.], Springer Science and Business Media LLC, v. 2018, n. 9, 2018.

18 BOER, W. de. *Precision Experiments at LEP*. In: *Advanced Series on Directions in High Energy Physics*. [S.l.: s.n.], 2015. p. 107–136.

19 CERN: EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH. *Large Hardon Collider*. European Organization for Nuclear Research. Disponível em: https://home.cern/science/accelerators/large-hadron-collider. Acesso em: 18 jan. 2023.

20 CERN: EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH. *CERN's accelerator complex*. European Organization for Nuclear Research. Disponível em: https://home.cern/science/accelerators/accelerator-complex. Acesso em: 18 jan. 2023.

21 ALDÁ JÚNIOR, W. L. *Estudo da produção de multijatos em colisões próton-próton com $\sqrt{s}$ = 7 TeV no detector CMS/LHC*. 2013. 123 p. Tese (Doutorado em Física) — Instituto de Física Armando Dias Tavares, Rio de Janeiro, 2013.

22 CMS COLLABORATION. Strategies and performance of the CMS silicon tracker alignment during LHC run 2. *Nuclear Instruments and Methods in Physics Research Section A*, [S.l.], v. 1037, p. 166795, 2022. ISSN 0168-9002.

23 SIRUNYAN, A. et al. *Particle-flow reconstruction and global event description with the CMS detector. Journal of Instrumentation*, [S.l.], IOP Publishing, v. 12, n. 10, p. P10003–P10003, 2017.

24 PEREZ RIVERA, G. M. *Unitarization Models For Vector Boson Scattering at the LHC*. 2018. 101 p. Thesis (PhD) — Fakultät für Physik, Karlsruher Instituts für Technologie, Karlsruhe, 2018.

25 NEUTELINGS, I. *CMS coordinate system*. Disponível em: https://tikz.net/axis3d_cms/. Acesso em: 18 jan. 2023.

26 BAUER, J. *Perspektiven zur Beobachtung der elektroschwachen Produktion einzelner Top-Quarks mit dem CMS Experiment*. [S.l.: s.n.]. 2010. P. 158.

27 XAVIER, F. M. V. *Recepção de sinal de múons no calorímetro hadrônico do experimento atlas*. 2011. 96 f. Dissertação (Mestrado em Engenharia Elétrica) — Faculdade de Engenharia, Universidade Federal de Juiz de Fora, Juiz de Fora, 2011.

28 BIALAS, W.; PETYT, D. A. *Mitigation of anomalous APD signals in the CMS ECAL. Journal of Instrumentation*, [S.l.], v. 8, n. 03, p. C03020, 2013.

29 BARTOLONI, A. et al. *The CMS ECAL barrel HV system. Journal of Instrumentation*, [S.l.], v. 8, p. C02039, 2013.

30 AVEZOV, A. et al. *The Hadron Calorimeter Project Technical Design Report*. [S.l.: s.n.]. 1997.

31 CHATRCHYAN, S. et al. *The CMS Experiment at the CERN LHC. JINST*, [S.l.], v. 3, p. S08004, 2008.

32 FOCARDI, E. *Status of the CMS Detector*. [S.l.: s.n.]. 2012.

33 CMS COLLABORATION. *Performance of the CMS hadron calorimeter with cosmic ray muons and LHC beam data. Journal of Instrumentation*, [S.l.], v. 5, p. T03012, 2010.

34 PAOLUCCI, P. *The CMS Muon system*. Geneva, n. CMS-CR-2006-006, 2005. Disponível em: https://cds.cern.ch/record/927394. Acesso em: 8 dez. 2022.

35 THYSSEN, F. *Commissioning, operation and performance of the CMS resistive plate chamber system*. 2014. 155 p. Tese (Doutorado) — Faculty of Sciences, Ghent University, Ghent, 2014.

36 CHAUHAN, S.; CHOUDHARY, B. *Search for Quark Compositeness at* $\sqrt{s}$ *= 14 TeV at the Large Hadron Collider*. [S.l.: s.n.]. 2009.

37 CMS COLLABORATION. *The CMS high level trigger. The European Physical Journal C*, [S.l.], Springer Science and Business Media LLC, v. 46, n. 3, p. 605–667, 2006.

38 SIRUNYAN, A. *Electron and photon reconstruction and identification with the CMS experiment at the CERN LHC. Journal of Instrumentation*, [S.l.], IOP Publishing, v. 16, n. 05, p. P05014, 2021.

39 CMS COLLABORATION. *Identification of b-quark jets with the CMS experiment. Journal of Instrumentation*, [S.l.], IOP Publishing, v. 8, n. 04, p. P04013–P04013, 2013.

40 FERRO, C. *B-tagging in CMS*. [S.l.], arXiv, 2012. Disponível em: https://arxiv.org/abs/1201.5292. Acesso em: 28 nov. 2022.

41 SIRUNYAN, A. *Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV. Journal of Instrumentation*, [S.l.], IOP Publishing, v. 13, n. 05, p. P05011–P05011, 2018.

42 BOLS, E. et al. *Jet flavour classification using DeepJet. Journal of Instrumentation*, [S.l.], IOP Publishing, v. 15, n. 12, p. P12012–P12012, 2020. Acesso em: 5 dez. 2022.

43 BISHOP, C. M. *Pattern Recognition and Machine Learning*. [S.l.]: Springer, 2006.

44 MYLES, A. et al. *An Introduction to Decision Tree Modeling. Journal of Chemometrics*, [S.l.], v. 18, 2004.

45 ADAMS, T. et al. *Gravitational-Wave Detection using Multivariate Analysis. Physical Review D*, [S.l.], v. 88, 2013.

46 WITTEN, I.; FRANK, E.; HALL, M. *An Introduction to Decision Trees. Data Mining and Knowledge Discovery*, [S.l.], v. 2, n. 4, 1998.

47 CHEN, T.; GUESTRIN, C. *XGBoost*: A scalable tree boosting system. *In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 22*, 2016. San Francisco. Proceedings... [San Francisco: ACM], 2016.

48 HE, Z. et al. *Gradient Boosting Machine*: A survey. [S.l.], arXiv, 2019. Disponível em: https://arxiv.org/abs/1908.06951. Acesso em: 12 jan. 2023.

49 ADAM-BOURDARIOS, C. et al. *The Higgs Machine Learning Challenge. Journal of Physics: Conference Series*, [S.l.], IOP Publishing, v. 664, n. 7, p. 072015, 2015.

50 ZHANG, T. et al. *Improving Convection Trigger Functions in Deep Convective Parameterization Schemes Using Machine Learning. Journal of Advances in Modeling Earth Systems*, [S.l.], v. 13, 2021.

51 ROSENBLATT, F. *The perceptron*: a probabilistic model for information storage and organization in the brain. *Psychological review*, [S.l.], American Psychological Association, v. 65, n. 6, p. 386, 1958.

52 GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016. Disponível em: http://www.deeplearningbook.org. Acesso em: 17 nov. 2022.

53 MOKHTAR, K. Z.; MOHAMAD-SALEH, J. *An Oil Fraction Neural Sensor Developed Using Electrical Capacitance Tomography Sensor Data. Sensors (Basel, Switzerland)*, [S.l.], v. 13, p. 11385–406, 2013.

54 CMS COLLABORATION. *The CMS NanoAOD data tier*. 2022. Disponível em: https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookNanoAOD. Acesso em: 10 jan. 2023.

55 REIS, M. C. *Estimativa da eficiência de triggers em análise de dados de busca por matéria escura no CMS,LHC*. 2023. Dissertação (Mestrado em Física) — Insituto de Física Armando Dias Tavares, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2023.

56 CMS COLLABORATION. lectron reconstruction and identification at sqrt(s) = 7 TeV. Geneva, 2010. Disponível em: https://cds.cern.ch/record/1299116. Acesso em: 2 dez. 2022.

57 SIRUNYAN, A. M. et al. *Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at sqrt(s) = 13 TeV. Journal of Instrumentation*, [S.l.], IOP Publishing, v. 13, n. 06, p. P06015–P06015, 2018.

58 CMS COLLABORATION. Jet algorithms performance in 13 tev data. 2017. Disponível em: https://cds.cern.ch/record/2256875. Acesso em: 5 dez. 2022.

59 CMS COLLABORATION. *Pileup Jet Identification*. Geneva, 2013. Disponível em: https://cds.cern.ch/record/1581583. Acesso em: 4 dez. 2022.

60 CMS COLLABORATION. *MET Filter Recommendations for Run II*. 2022a. Disponível em: https://twiki.cern.ch/twiki/bin/viewauth/CMS/MissingETOptionalFiltersRun2. Acesso em: 10 jan. 2023.

61 PÉREZ ADÁN, D. *Comparação da energia transversa perdida dos pontos de sinal propostos*. 2021. Figura desenhada exclusivamente para essa pesquisa.

62 CMS COLLABORATION. *Jet Energy Corrections*: Official software tools for applying jec corrections and uncertainties. 2020. Disponível em: https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookJetEnergyCorrections. Acesso em: 10 jan. 2023.

63 CMS COLLABORATION. *Jet Energy Resolution*. 2023. Disponível em: https://twiki.cern.ch/twiki/bin/view/CMS/JetResolution. Acesso em: 10 jan. 2023.

64 CMS COLLABORATION. *B-Tagging*. 2022b. Disponível em: https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookBTagging. Acesso em: 10 jan. 2023.

65 CMS COLLABORATION. *MET Analysis*. 2019. Disponível em: https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookMetAnalysis. Acesso em: 10 jan. 2023.

66 CMS COLLABORATION. *Rochester corrections*. 2021a. Disponível em: https://twiki.cern.ch/twiki/bin/viewauth/CMS/RochcorMuon. Acesso em: 10 jan. 2023.

67 CMS COLLABORATION. *Reweighting recipe to emulate Level 1 ECAL prefiring*. 2021b. Disponível em: https://twiki.cern.ch/twiki/bin/viewauth/CMS/L1ECALPrefiringWeightRecipe. Acesso em: 10 jan. 2023.

68 CMS COLLABORATION. *Jet identification in high pile-up environment (PileupJetID) for Ultra Legacy Data*. 2022c. Disponível em: https://twiki.cern.ch/twiki/bin/view/CMS/PileupJetIDUL. Acesso em: 10 jan. 2023.

69 CMS COLLABORATION. *Methods to apply b-tagging efficiency scale factors*. 2022d. Disponível em: https://twiki.cern.ch/twiki/bin/viewauth/CMS/BTagSFMethods. Acesso em: 10 jan. 2023.

70 CMS COLLABORATION. *Jet Flavour Identification (MC Truth)*. 2016. Disponível em: https://twiki.cern.ch/twiki/bin/view/CMSPublic/SWGuideBTagMCTools. Acesso em: 10 jan. 2023.

71 LIGT, K. J. *Effects of model parameter variations for the $\bar{b}bH \rightarrow Za \rightarrow \ell\bar{\ell}\chi\bar{\chi}$ process in the performance of multivariate classifiers*. 2022. 81 p. Dissertação (Mestrado em Física) — Universade de Hamburgo, Hamburgo, 2022.

72 CHOLLET, F. et al. *Keras*. 2015. Disponível em: https://keras.io. Acesso em: 13 dez. 2022.

73 ABADI, M. et al. *TensorFlow*: Large-scale machine learning on heterogeneous systems. 2015. Software available from tensorflow.org. Disponível em: https://www.tensorflow.org/. Acesso em: 13 dez. 2022.

74 VALASSI, A. *Binary classifier metrics for optimizing HEP event selection. EPJ Web of Conferences*, [S.l.], v. 214, p. 06004, 2019.

75 BISHOP, C. *Neural networks for pattern recognition*. [S.l.]: Oxford University Press, USA, 1995.

76 CAMPOS, G. M. S. *hhdmAnalysis-mestrado*: Base de código utilizada para produção de modelos de ML e figuras para pesquisa de mestrado. 2023. Disponível em: https://github.com/gabrielmscampos/hhdmAnalysis-mestrado. Acesso em: 16 mar. 2023.

77 LESTER, C.; NACHMAN, B. *Bisection-based asymmetric M T2 computation*: a higher precision calculator than existing symmetric methods. *Journal of High Energy Physics*, [S.l.], v. 2015, 2015.

78 CMS COLLABORATION. *AlCaTools*: Collection of tools for daily work of cms alca/db team. 2023. Disponível em: https://github.com/cms-AlCaDB/AlCaTools. Acesso em: 12 nov. 2021.

APÊNDICE A – Amostras de dados

Tabela 5 - Amostras de dados de todos os períodos utilizados na análise

Datasets	Eras	Nome	$\mathcal{L}_{int}$ (fb^{-1})
		2016 pre-VFP	
DoubleEG	B, C, D, E, F	HIPM_UL2016_MiniAODv2_NanoAODv9v2	19,5
DoubleMuon		ver1_HIPM_UL2016_MiniAODv2_NanoAODv9v2	
MuonEG		ver2_HIPM_UL2016_MiniAODv2_NanoAODv9v2	
SingleElectron		ver2_HIPM_UL2016_MiniAODv2_NanoAODv9v3	
SingleMuon			
		2016 post-VFP	
DoubleEG	F, G, H	UL2016_MiniAODv2_NanoAODv9v1	16,8
DoubleMuon		UL2016_MiniAODv2_NanoAODv9v2	
MuonEG			
SingleElectron			
SingleMuon			
		2017	
DoubleEG	B, C, D, E, F	UL2017_MiniAODv2_NanoAODv9v1	41,4
DoubleMuon			
MuonEG			
SingleElectron			
SingleMuon			
		2018	
DoubleMuon	A, B, C, D	UL2018_MiniAODv2_NanoAODv9v1	59,8
EGamma		UL2018_MiniAODv2_NanoAODv9v2	
MuonEG		UL2018_MiniAODv2_NanoAODv9v3	
SingleMuon			

Fonte: O autor, 2023.

APÊNDICE B – Amostras de simulação

Amostras de simulação produzidas pela Colaboração CMS. As abreviações "ext" descritas na lista abaixo serão utilizadas para simplificar a escrita dos nomes dos conjuntos. A coluna σ é a seção de choque, a coluna L é a luminosidade e a coluna "Acurácia" descreve a ordem de aproximação utilizada na definição matemática do processo físico, *Leading Order* (LO), *Next Leading Order* (NLO) e *unknown* (quando a acurácia é desconhecida).

- ext1: TuneCP5_13TeV_madgraph-pythia8
- ext2: TuneCP5_13TeV-amcatnloFXFX-pythia8
- ext3: TuneCP5_13TeV-powheg-pythia8
- ext4: TuneCP5_13TeV-pythia8
- ext5: TuneCP5_13TeV-amcatnlo-pythia8
- ext6: TuneCP5_13TeV-amcatnloFXFX-madspin-pythia8
- ext7: TuneCP5_13TeV_powheg2_JHUGenV714_pythia8
- ext8: TuneCP5_13TeV_powheg2_JHUGenV7011_pythia8
- ext9: TuneCP5_13TeV_powheg2-minlo-HWJ_JHUGenV7011_pythia8
- ext10: TuneCP5_13TeV_powheg2-minlo-HZJ_JHUGenV7011_pythia8
- ext11: TuneCP5_13TeV-madgraphMLM-pythia8

Tabela 6 - Amostras de Sinal (Signal_M_H_M_a)

Nome da amostra	$\sigma[pb]$	L $[fb^{-1}]$	Acurácia
bbHToZaToLLChiChi_2HDMa_MH-400_Ma-100_MChi-45_[ext1]	0.0449300	22260.0	NLO
bbHToZaToLLChiChi_2HDMa_MH-400_Ma-200_MChi-45_[ext1]	0.0339000	29500.0	NLO
bbHToZaToLLChiChi_2HDMa_MH-500_Ma-100_MChi-45_[ext1]	0.0208900	47870.0	NLO
bbHToZaToLLChiChi_2HDMa_MH-500_Ma-200_MChi-45_[ext1]	0.0193100	51780.0	NLO
bbHToZaToLLChiChi_2HDMa_MH-500_Ma-300_MChi-45_[ext1]	0.0127500	78430.0	NLO
bbHToZaToLLChiChi_2HDMa_MH-600_Ma-100_MChi-45_[ext1]	0.0099530	100500.0	NLO
bbHToZaToLLChiChi_2HDMa_MH-600_Ma-200_MChi-45_[ext1]	0.0098490	101500.0	NLO
bbHToZaToLLChiChi_2HDMa_MH-600_Ma-300_MChi-45_[ext1]	0.0086210	116000.0	NLO
bbHToZaToLLChiChi_2HDMa_MH-600_Ma-400_MChi-45_[ext1]	0.0052960	188800.0	NLO
bbHToZaToLLChiChi_2HDMa_MH-800_Ma-100_MChi-45_[ext1]	0.0025990	384800.0	NLO
bbHToZaToLLChiChi_2HDMa_MH-800_Ma-200_MChi-45_[ext1]	0.0026800	373200.0	NLO
bbHToZaToLLChiChi_2HDMa_MH-800_Ma-300_MChi-45_[ext1]	0.0025800	387600.0	NLO
bbHToZaToLLChiChi_2HDMa_MH-800_Ma-400_MChi-45_[ext1]	0.0024130	414500.0	NLO
bbHToZaToLLChiChi_2HDMa_MH-800_Ma-600_MChi-45_[ext1]	0.0011860	843000.0	NLO
bbHToZaToLLChiChi_2HDMa_MH-1000_Ma-100_MChi-45_[ext1]	0.0008098	1235000.0	NLO
bbHToZaToLLChiChi_2HDMa_MH-1000_Ma-200_MChi-45_[ext1]	0.0008386	1192000.0	NLO
bbHToZaToLLChiChi_2HDMa_MH-1000_Ma-300_MChi-45_[ext1]	0.0008243	1213000.0	NLO
bbHToZaToLLChiChi_2HDMa_MH-1000_Ma-400_MChi-45_[ext1]	0.0008025	1246000.0	NLO
bbHToZaToLLChiChi_2HDMa_MH-1000_Ma-600_MChi-45_[ext1]	0.0007163	1396000.0	NLO
bbHToZaToLLChiChi_2HDMa_MH-1000_Ma-800_MChi-45_[ext1]	0.0003395	2945000.0	NLO

Fonte: O autor, 2023.

Tabela 7 - Amostras de Drell-Yan + Jatos (DYJetsToLL)

Nome da amostra	$\sigma[pb]$	L $[fb^{-1}]$	Acurácia
DYJetsToLL_M-50_[ext2]	6404.0000	0.07039	unknown
DYJetsToLL_LHEFilterPtZ-50To100_MatchEWPDG20_[ext2]	397.4000	0.42910	unknown
DYJetsToLL_LHEFilterPtZ-100To250_MatchEWPDG20_[ext2]	97.2000	1.69600	unknown
DYJetsToLL_LHEFilterPtZ-250To400_MatchEWPDG20_[ext2]	3.70100	48.11000	unknown
DYJetsToLL_LHEFilterPtZ-400To650_MatchEWPDG20_[ext2]	0.50860	383.20000	unknown
DYJetsToLL_LHEFilterPtZ-650ToInf_MatchEWPDG20_[ext2]	0.04728	4538.00000	unknown

Fonte: O autor, 2023.

Tabela 8 - Amostras de pares de quark top ($t\bar{t}$)

Nome da amostra	$\sigma[pb]$	L $[fb^{-1}]$	Acurácia
TTTo2L2Nu_[ext3]	687.1	1.433	NLO
TTToSemiLeptonic_[ext3]	687.1	1.433	NLO

Fonte: O autor, 2023.

Tabela 9 - Amostras de quark top (ST)

Nome da amostra	$\sigma[pb]$	L $[fb^{-1}]$	Acurácia
ST_tW_top_5f_NoFullyHadronicDecays_[ext3]	32.45	30.810	NLO
ST_tW_antitop_5f_NoFullyHadronicDecays_[ext3]	32.51	30.760	NLO
ST_t-channel_top_5f_InclusiveDecays_[ext3]	119.70	8.265	NLO
ST_t-channel_antitop_5f_InclusiveDecays_[ext3]	71.74	13.800	NLO

Fonte: O autor, 2023.

Tabela 10 - Amostras de ZZ

Nome da amostra	$\sigma[pb]$	L $[fb^{-1}]$	Acurácia
ZZTo2L2Nu_[ext3]	0.9738	1022.0	NLO
ZZTo4L_[ext3]	1.3250	739.9	NLO

Fonte: O autor, 2023.

Tabela 11 - Amostras de WZ

Nome da amostra	$\sigma[pb]$	L $[fb^{-1}]$	Acurácia
WZTo3LNu_mllmin4p0_[ext3]	4.664	210.1	NLO

Fonte: O autor, 2023.

Tabela 12 - Amostras Residuais

Nome da amostra	$\sigma[pb]$	L $[fb^{-1}]$	Acurácia
ZZ_[ext4]	12.170000	82.19000	unknown
WZ_[ext4]	27.590000	36.25000	unknown
WZTo2Q2L_mllmin4p0_[ext2]	6.41900	60.52000	unknown
WW_[ext4]	75.950000	13.17000	unknown
WWTo2L2Nu_[ext3]	11.090000	89.45000	NLO
ZZZ_[ext5]	0.014760	54300.00000	unknown
WZZ_[ext5]	0.057090	14460.00000	unknown
WWZ_4F_[ext5]	0.170700	4869.00000	unknown
WWW_4F_[ext5]	0.215800	3786.00000	unknown
TTZZ_[ext1]	0.001386	716700.00000	LO
TTWW_[ext1]	0.007003	142800.00000	LO
TWZToLL_thad_Wlept_5f_DR_[ext5]	0.003004	116500.00000	unknown
TWZToLL_tlept_Whad_5f_DR_[ext5]	0.003004	117700.00000	unknown
TWZToLL_tlept_Wlept_5f_DR_[ext5]	0.001501	235400.00000	unknown
TTWJetsToLNu_[ext6]	0.216100	1360.00000	unknown
TTWJetsToQQ_[ext6]	0.437700	681.10000	unknown
TTZToQQ_[ext5]	0.510400	434.50000	unknown
TTZToNuNu_[ext5]	0.147600	1686.00000	unknown
TTZToLL_M-1to10_[ext5]	0.053240	5182.00000	unknown
tZq_ll_4f_ckm_NLO_[ext5]	0.075610	940.00000	unknown
ttHTobb_M125_[ext3]	0.526900	1821.00000	NLO
ttHToTauTau_M125_[ext3]	0.526900	1818.00000	NLO
GluGluHToWWTo2L2Nu_M125_[ext7]	21.470000	46.57000	NLO
GluGluHToZZTo4L_M125_[ext8]	28.870000	33.94000	NLO
WplusH_HToZZTo4L_M125_[ext9]	0.864800	1034.00000	NLO
WminusH_HToZZTo4L_M125_[ext9]	0.540900	1667.00000	NLO
ZH_HToZZ_4LFilter_M125_[ext10]	0.793500	1124.00000	NLO
WJetsToLNu_[ext11]	53870.000000	0.01856	LO
WGToLNuG_[ext11]	412.700000	2.42300	LO

Fonte: O autor, 2023.

APÊNDICE C – Lista completa de *triggers*

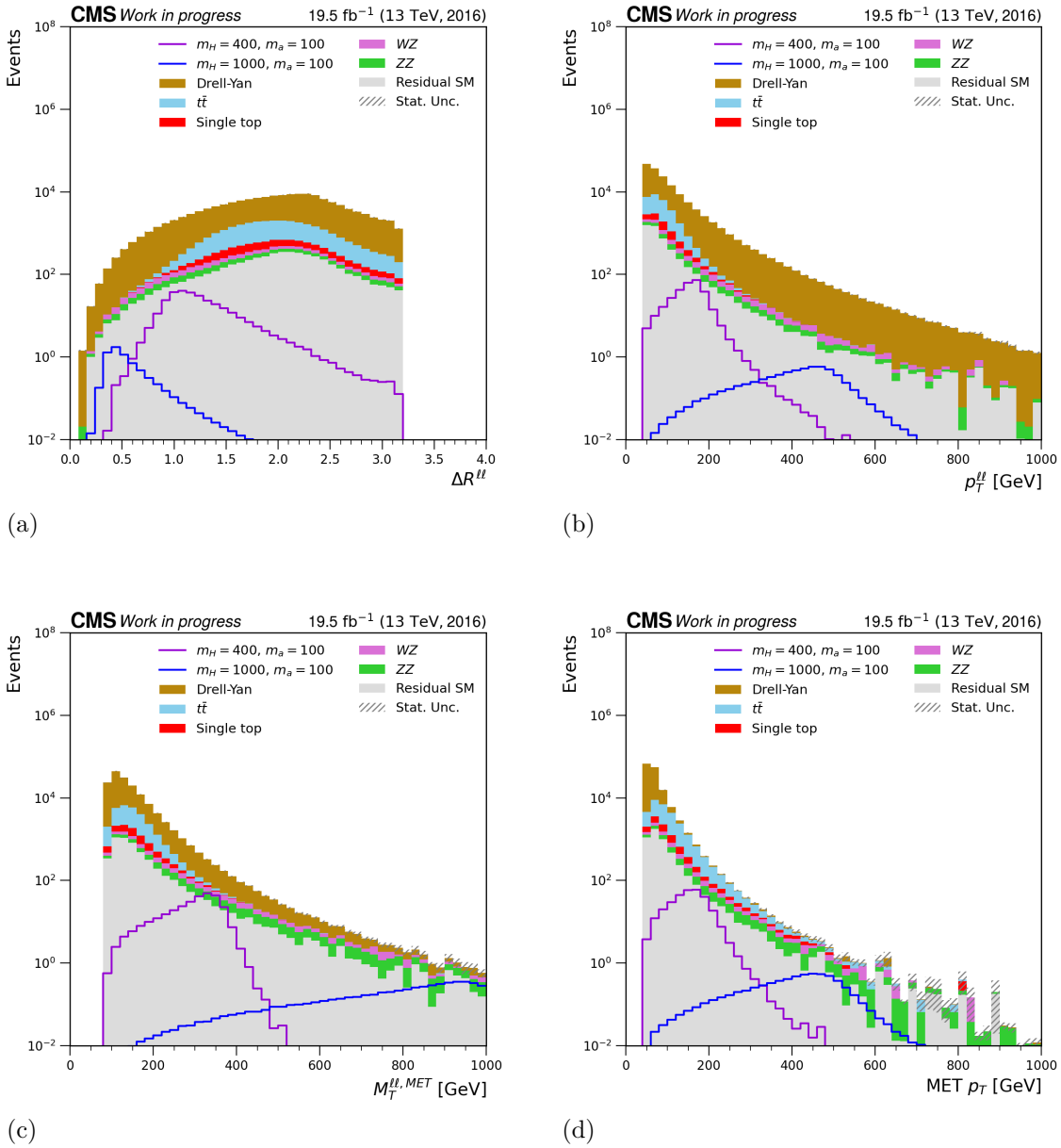
- 2016
  - HLT_Ele27_WPTight_Gsf
  - HLT_Ele115_CaloIdVT_GsfTrkIdT
  - HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL_DZ
  - HLT_IsoMu24
  - HLT_IsoTkMu24
  - HLT_Mu50
  - HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL(_DZ)
  - HLT_Mu17_TrkIsoVVL_TkMu8_TrkIsoVVL(_DZ)
  - HLT_Mu8_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL(_DZ)
  - HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL(_DZ)
- 2017
  - HLT_Ele35_WPTight_Gsf
  - HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL
  - HLT_DoubleEle33_CaloIdL_MW
  - HLT_IsoMu27
  - HLT_Mu50
  - HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_Mass8
  - HLT_Mu8_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL_DZ
  - HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVLHLT_Mu23_ TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL_DZ
- 2018
  - HLT_Ele32_WPTight_Gsf
  - HLT_Ele115_CaloIdVT_GsfTrkIdT
  - HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL
  - HLT_IsoMu24
  - HLT_Mu50
  - HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_Mass3p8

  - HLT_Mu8_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL_DZ
  - HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL

- $\not{E}_T$ *Triggers*
  - HLT_PFMET300
  - HLT_MET200
  - HLT_PFHT300_PFMET110
  - HLT_PFMET170_HBHECleaned
  - HLT_PFMET120_PFMHT120_IDTight
  - HLT_PFMETNoMu120_PFMHTNoMu120_IDTight

APÊNDICE D – Seleção de base nos demais períodos

Figura 35 - ΔR do sistema de dois léptons, p_T do sistema de dois léptons, massa transversa do sistema de dois léptons e $\not{E}_T$ e $\not{E}_T$ na seleção base para o período de 2016 pre-VFP

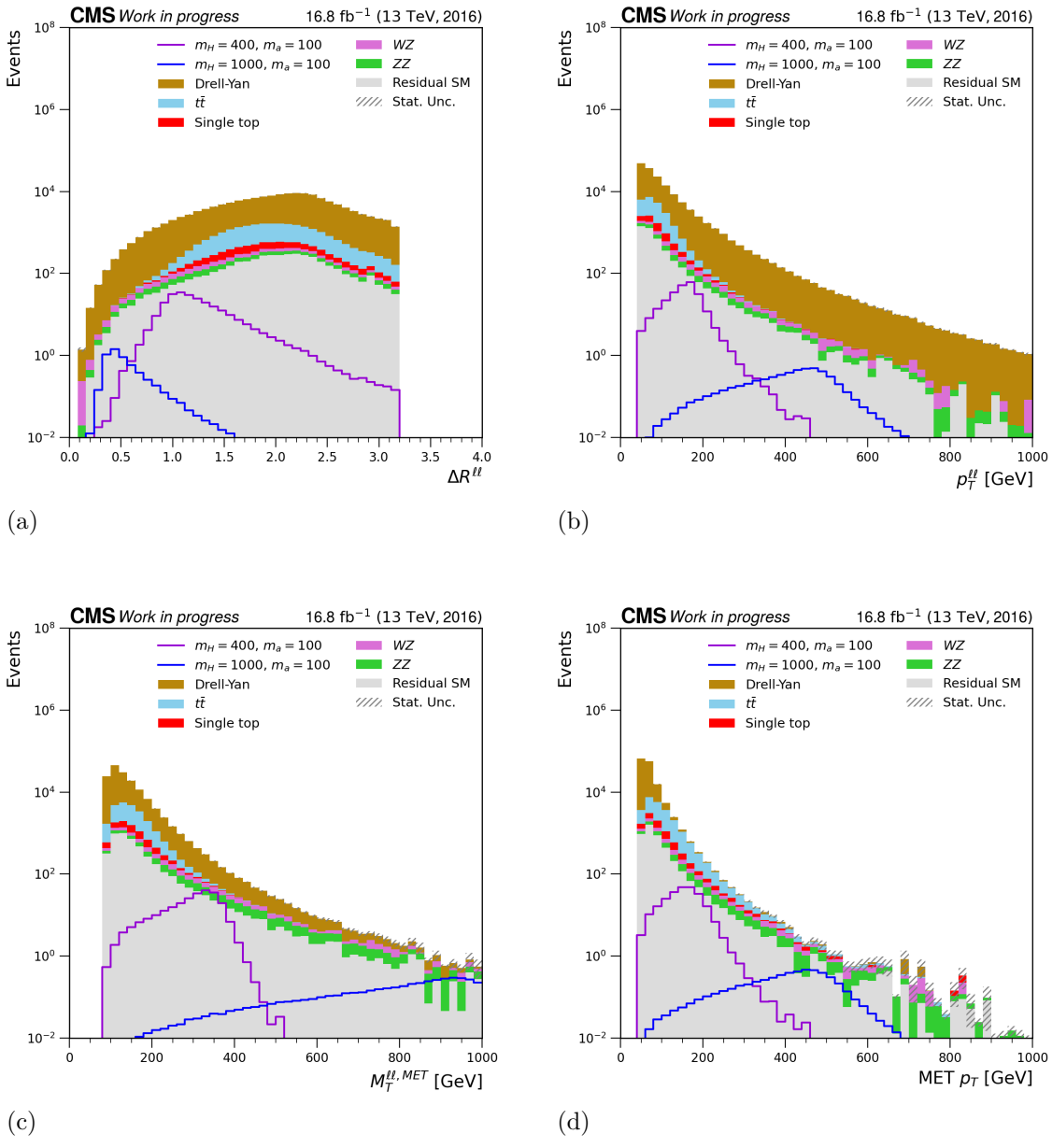


(a) (b)

(c) (d)

Fonte: O autor, 2023.

Figura 36 - ΔR do sistema de dois léptons, p_T do sistema de dois léptons, massa transversa do sistema de dois léptons e $\not{E}_T$ e $\not{E}_T$ na seleção base para o período de 2016 post-VFP

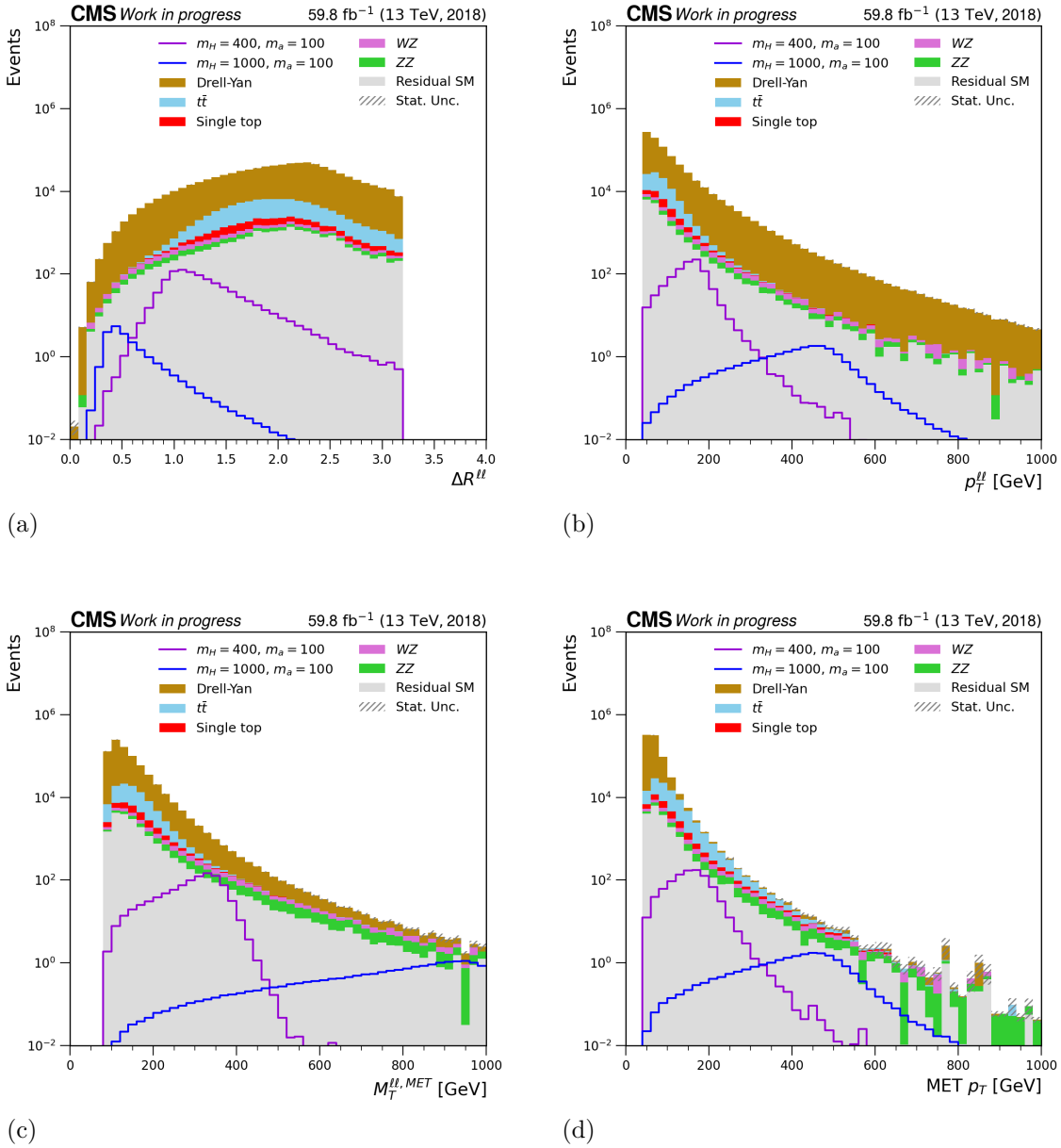


(a) (b)

(c) (d)

Fonte: O autor, 2023.

Figura 37 - ΔR do sistema de dois léptons, p_T do sistema de dois léptons, massa transversa do sistema de dois léptons e $\not{E}_T$ e $\not{E}_T$ na seleção base para o período de 2017



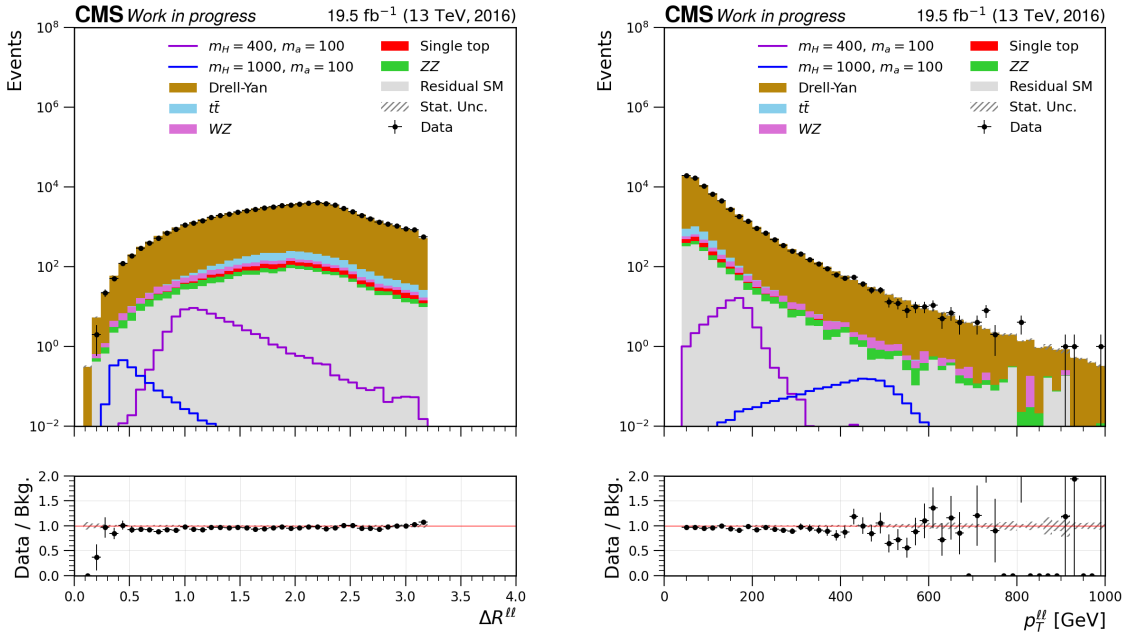
(a) (b)

(c) (d)

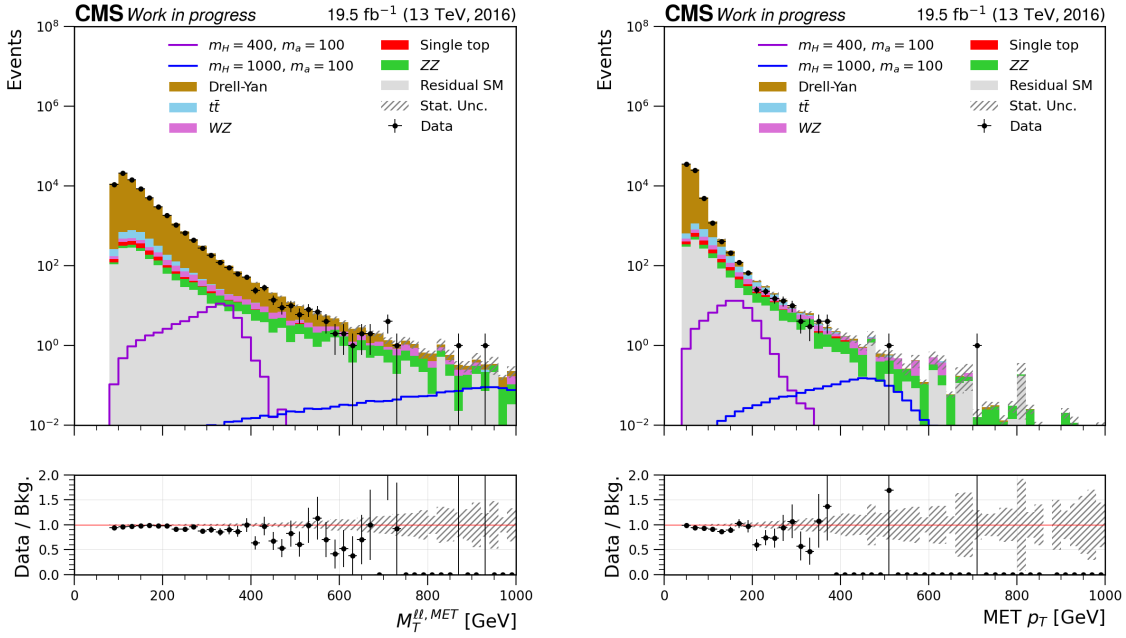
Fonte: O autor, 2023.

APÊNDICE E – Região de controle do DY nos demais períodos

Figura 38 - ΔR do sistema de dois léptons, p_T do sistema de dois léptons, massa transversa do sistema de dois léptons e $\not{E}_T$ e $\not{E}_T$ na região de controle do DY para o período de 2016 pre-VFP



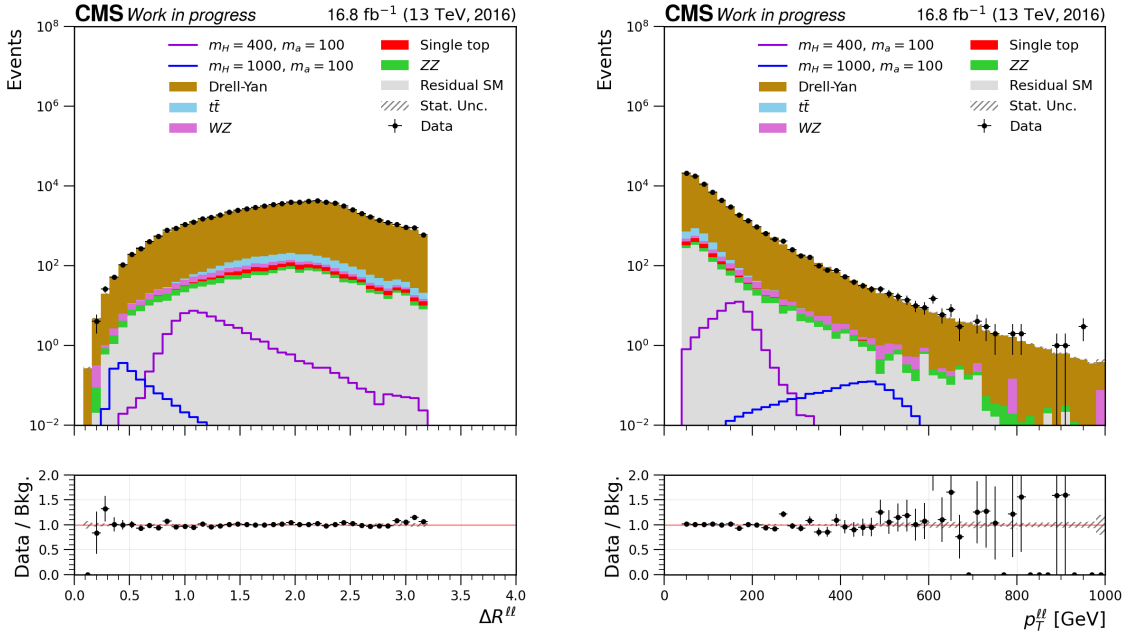
(a) (b)



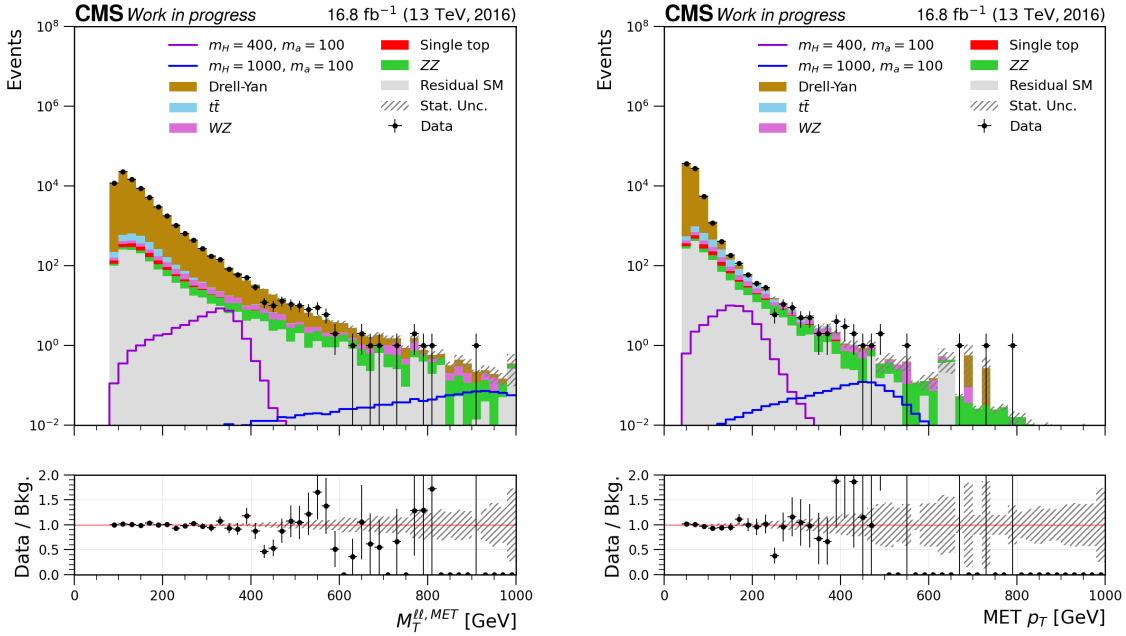
(c) (d)

Fonte: O autor, 2023.

Figura 39 - ΔR do sistema de dois léptons, p_T do sistema de dois léptons, massa transversa do sistema de dois léptons e $\not{E}_T$ e $\not{E}_T$ na região de controle do DY para o período de 2016 post-VFP



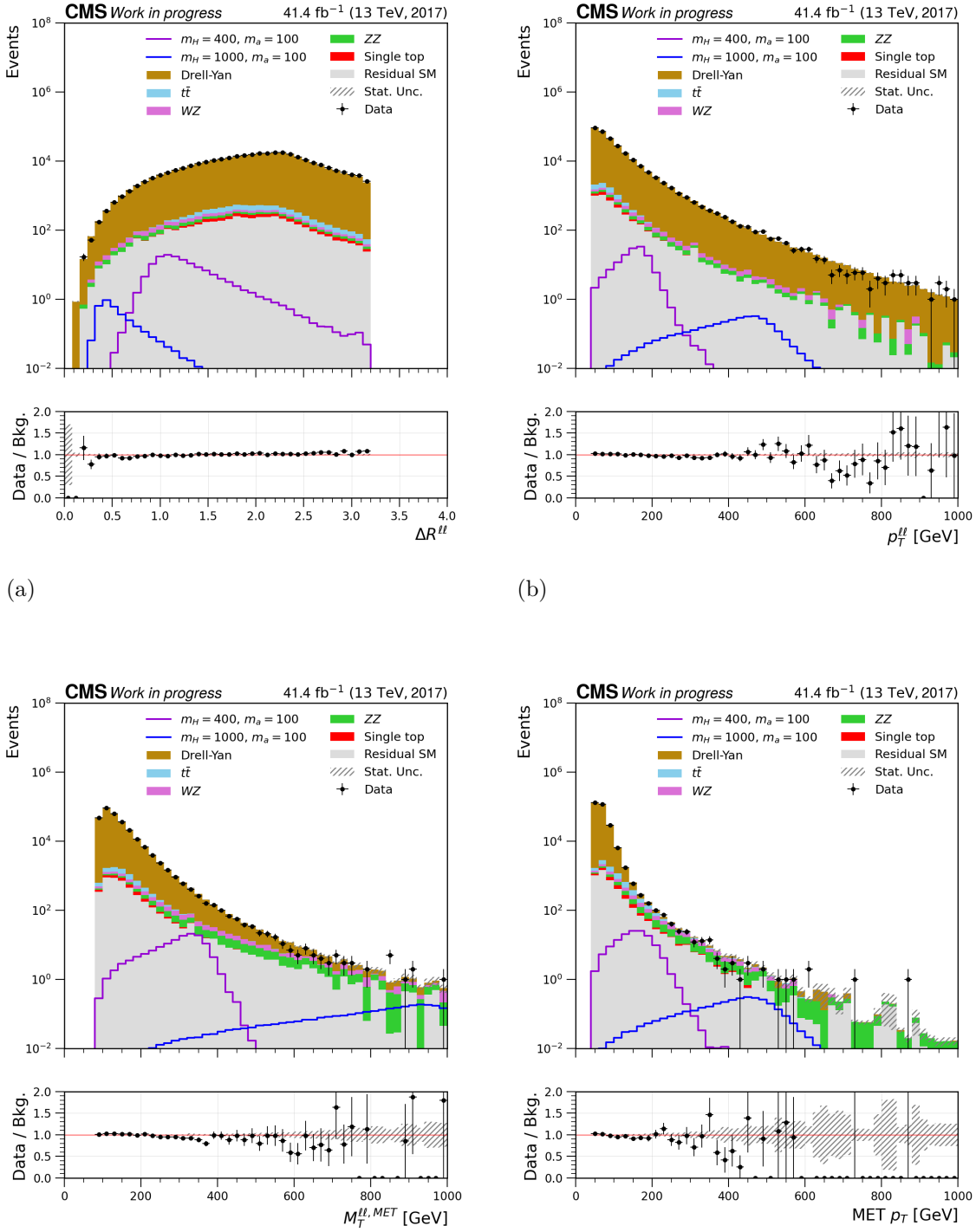
(a) (b)



(c) (d)

Fonte: O autor, 2023.

Figura 40 - ΔR do sistema de dois léptons, p_T do sistema de dois léptons, massa transversa do sistema de dois léptons e $\not{E}_T$ e $\not{E}_T$ na região de controle do DY para o período de 2017



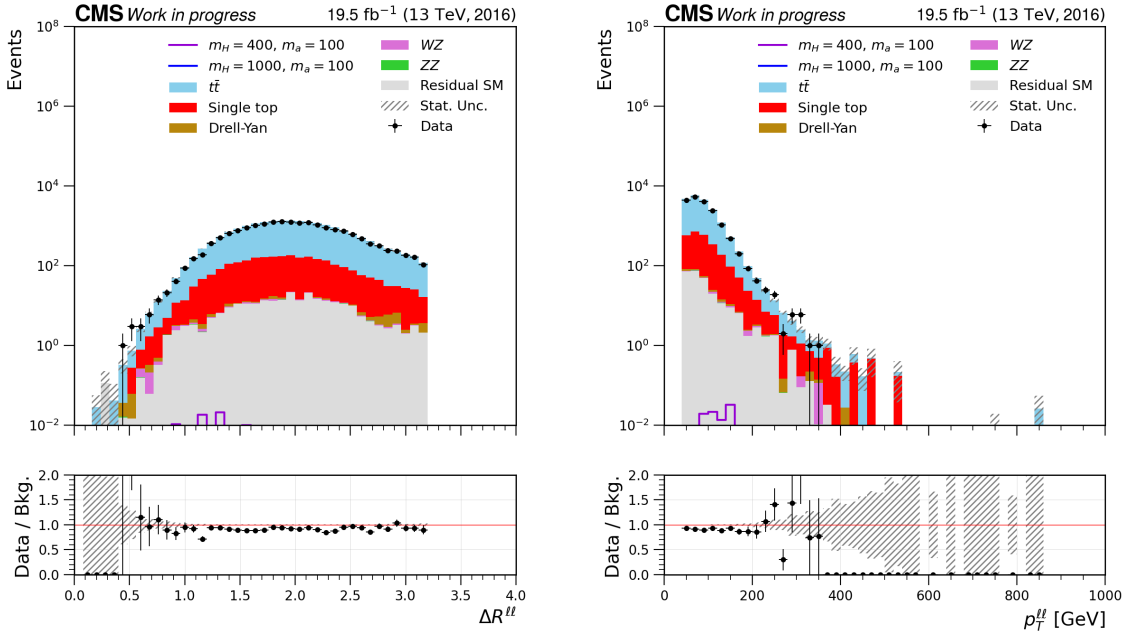
(a) (b)

(c) (d)

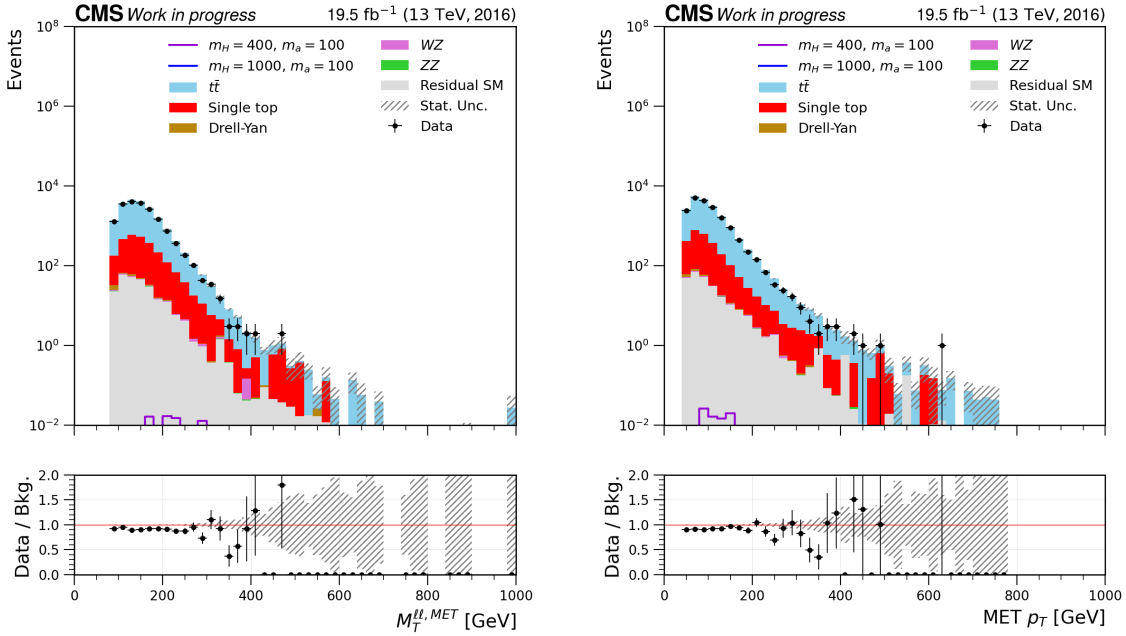
Fonte: O autor, 2023.

APÊNDICE F – Região de controle do $t\bar{t}$ nos demais períodos

Figura 41 - ΔR do sistema de dois léptons, p_T do sistema de dois léptons, massa transversa do sistema de dois léptons e $\not{E}_T$ e $\not{E}_T$ na região de controle do $t\bar{t}$ para o período de 2016 pre-VFP



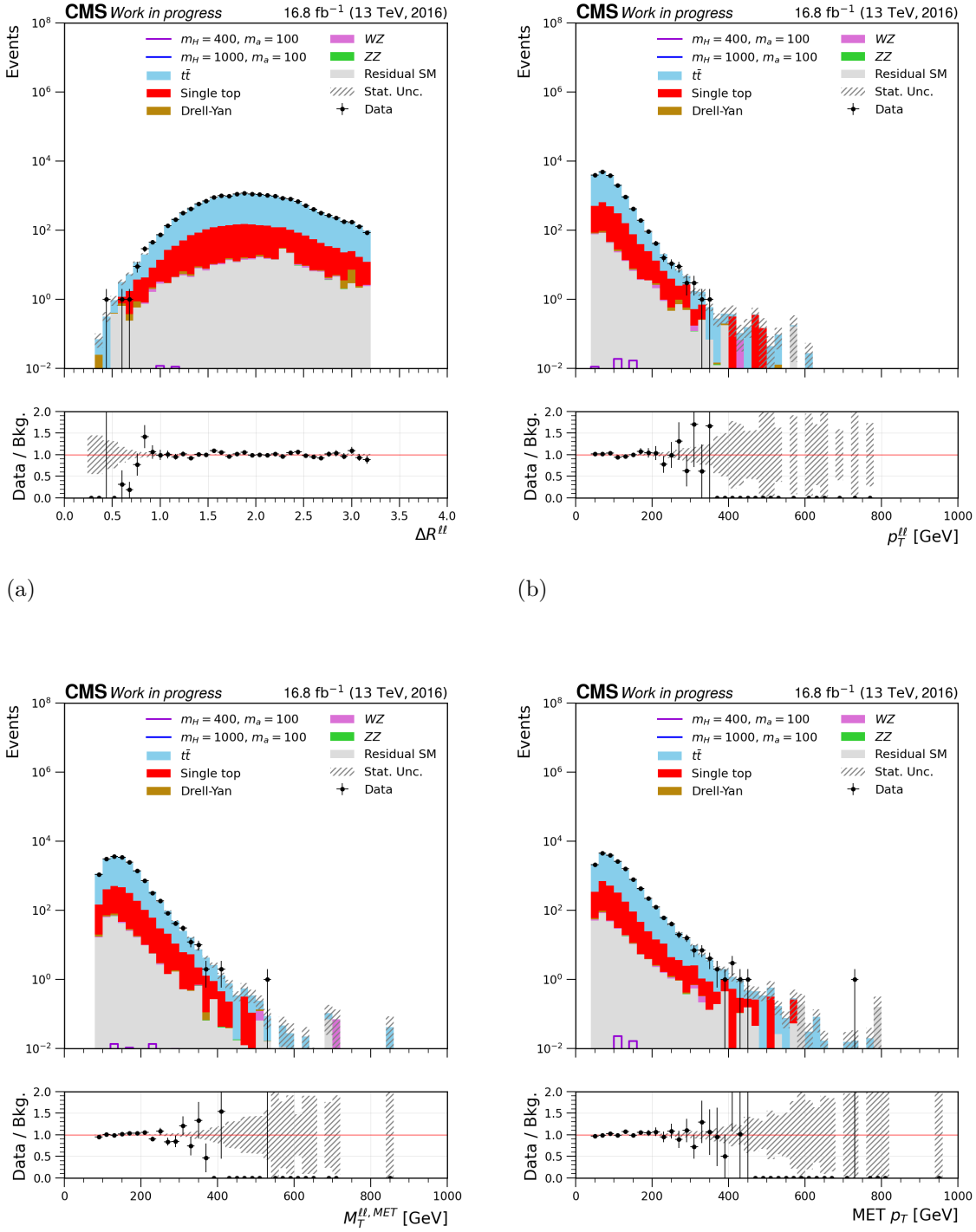
(a) (b)



(c) (d)

Fonte: O autor, 2023.

Figura 42 - ΔR do sistema de dois léptons, p_T do sistema de dois léptons, massa transversa do sistema de dois léptons e $\not{E}_T$ e $\not{E}_T$ na região de controle do $t\bar{t}$ para o período de 2016 post-VFP

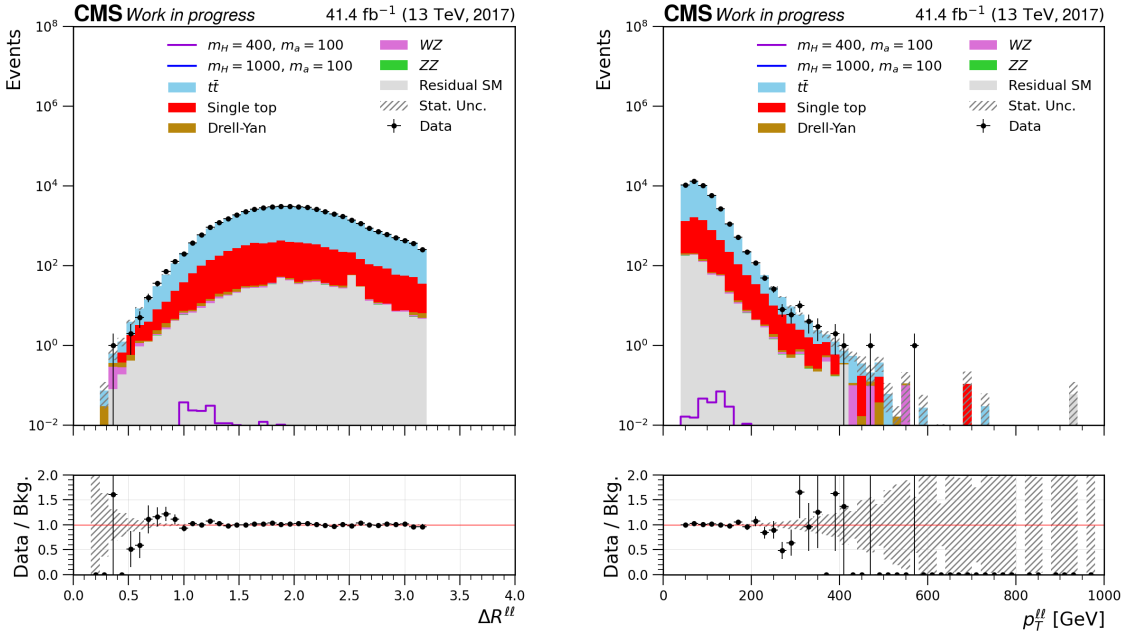


(a) (b)

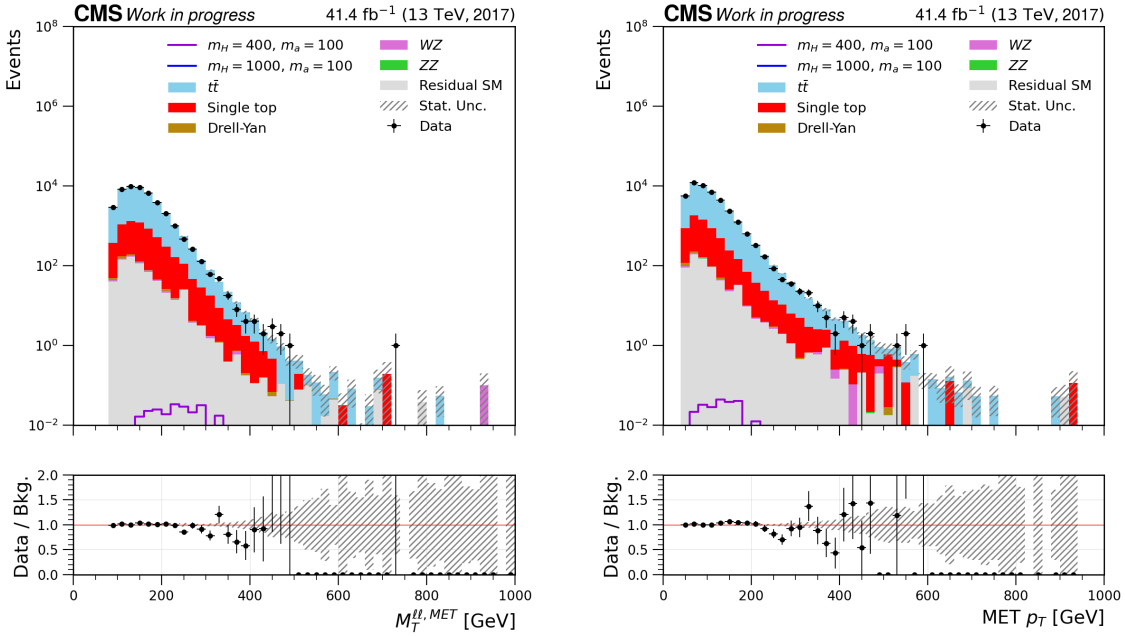
(c) (d)

Fonte: O autor, 2023.

Figura 43 - ΔR do sistema de dois léptons, p_T do sistema de dois léptons, massa transversa do sistema de dois léptons e $\not{E}_T$ e $\not{E}_T$ na região de controle do $t\bar{t}$ para o período de 2017



(a) (b)

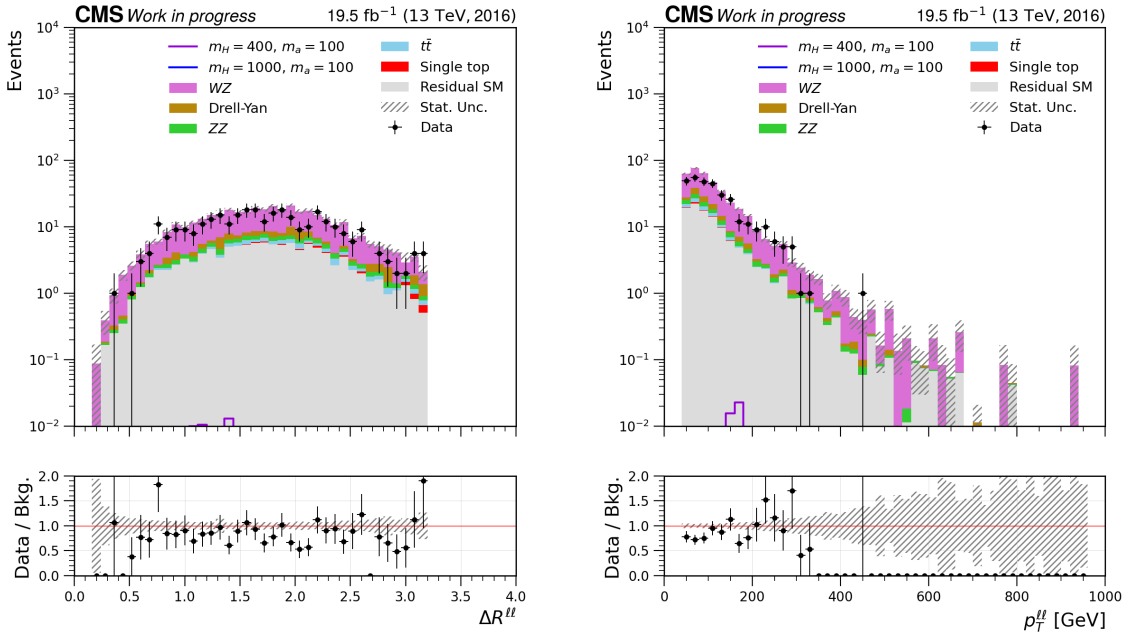


(c) (d)

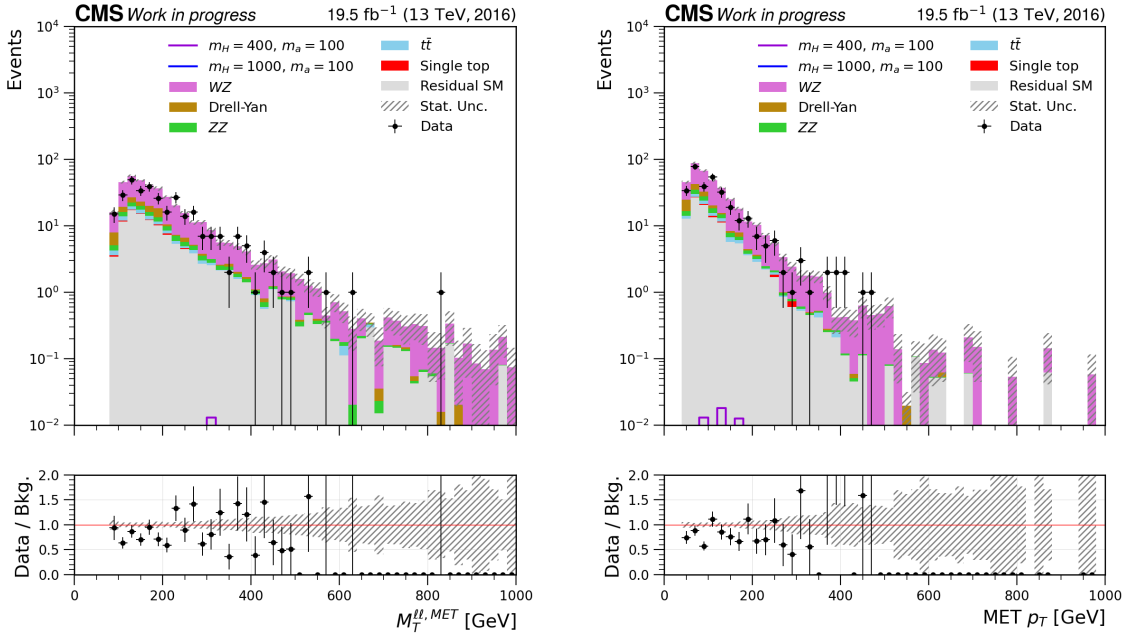
Fonte: O autor, 2023.

APÊNDICE G – Região de controle do WZ nos demais períodos

Figura 44 - ΔR do sistema de dois léptons, p_T do sistema de dois léptons, massa transversa do sistema de dois léptons e $\not{E}_T$ e $\not{E}_T$ na região de controle do WZ para o período de 2016 pre-VFP



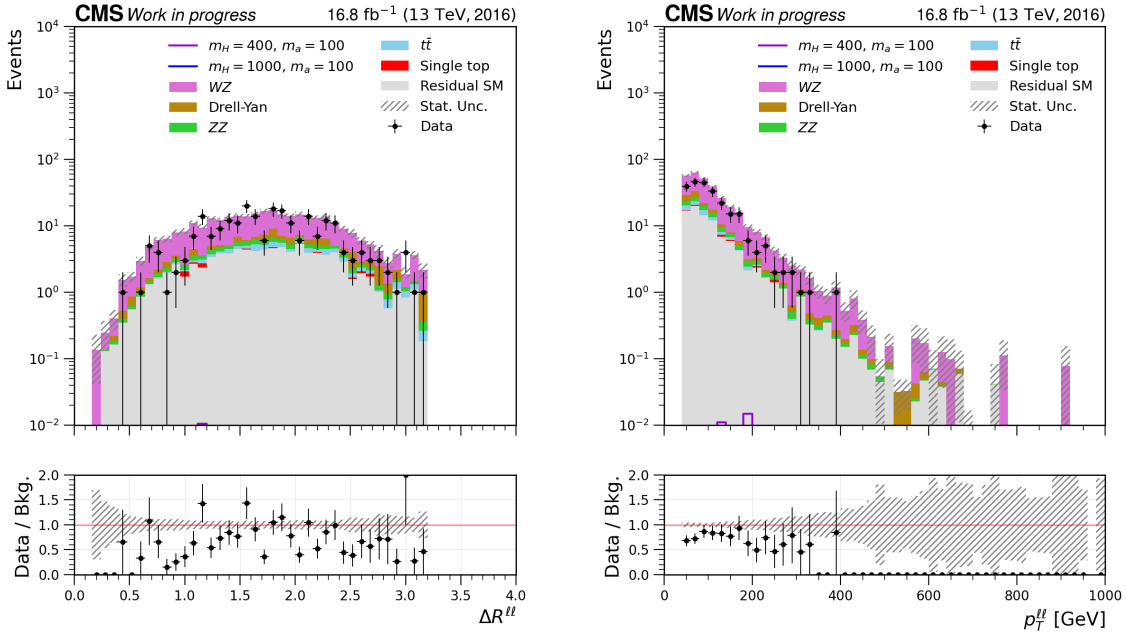
(a) (b)



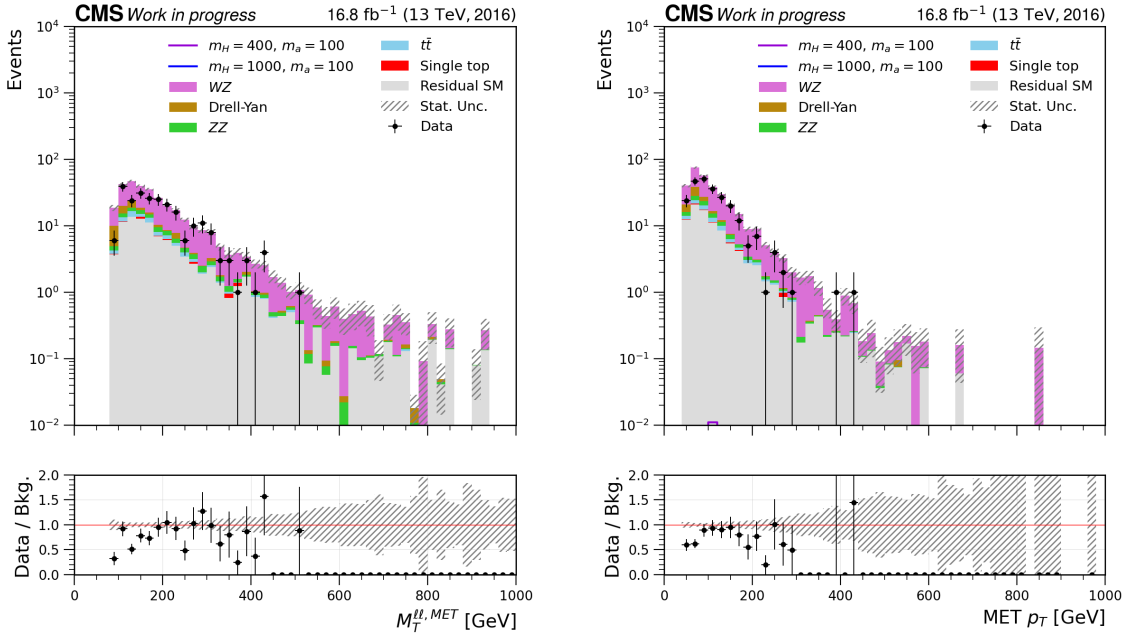
(c) (d)

Fonte: O autor, 2023.

Figura 45 - ΔR do sistema de dois léptons, p_T do sistema de dois léptons, massa transversa do sistema de dois léptons e $\not{E}_T$ e $\not{E}_T$ na região de controle do WZ para o período de 2016 post-VFP



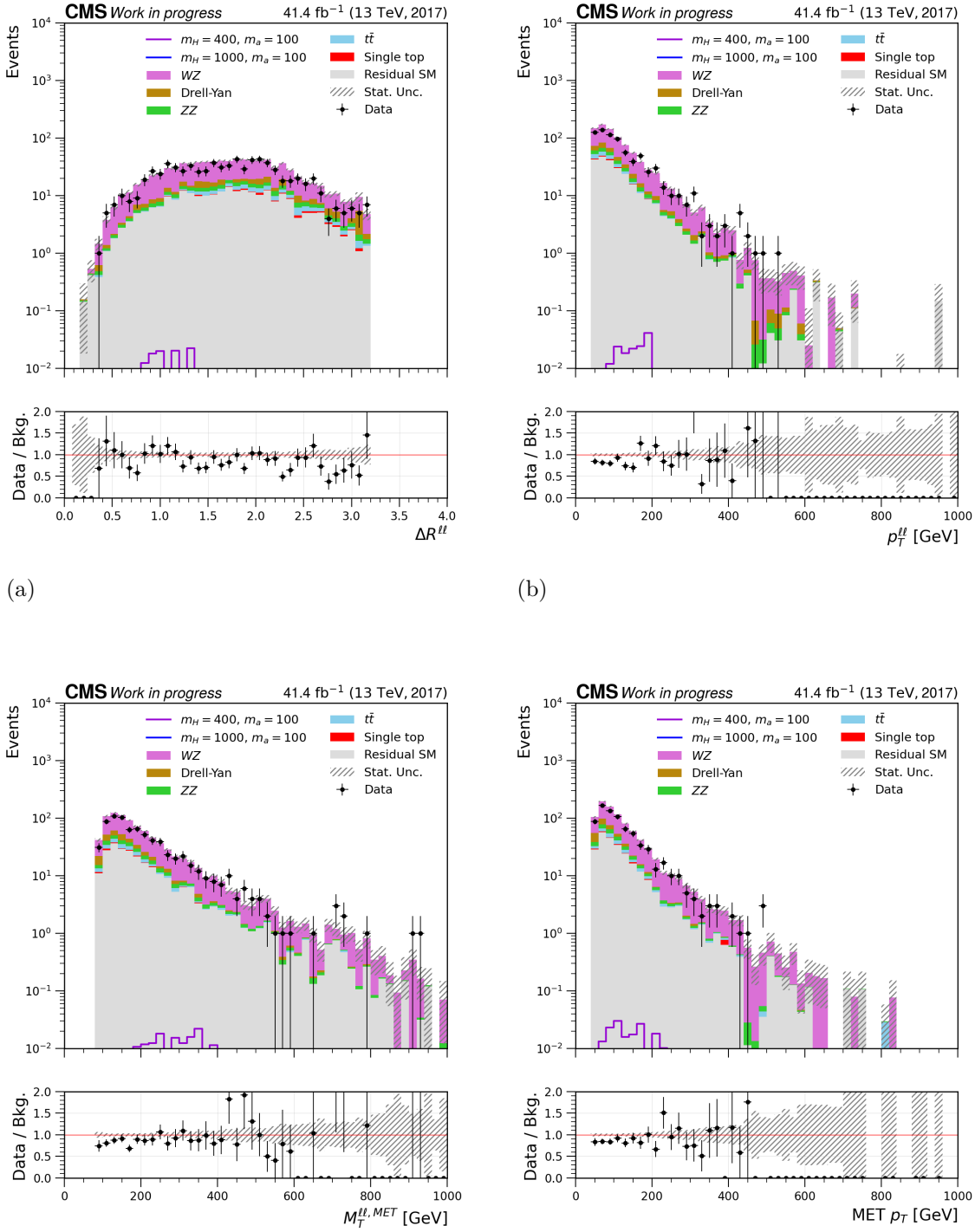
(a) (b)



(c) (d)

Fonte: O autor, 2023.

Figura 46 - ΔR do sistema de dois léptons, p_T do sistema de dois léptons, massa transversa do sistema de dois léptons e $\not{E}_T$ e $\not{E}_T$ na região de controle do WZ para o período de 2017



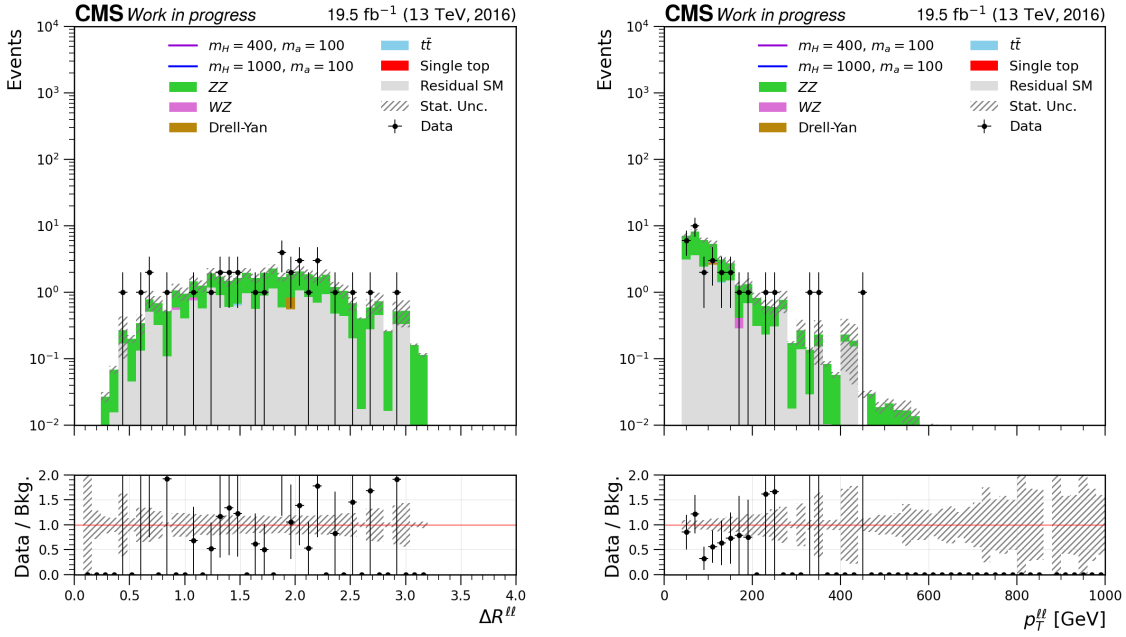
(a) (b)

(c) (d)

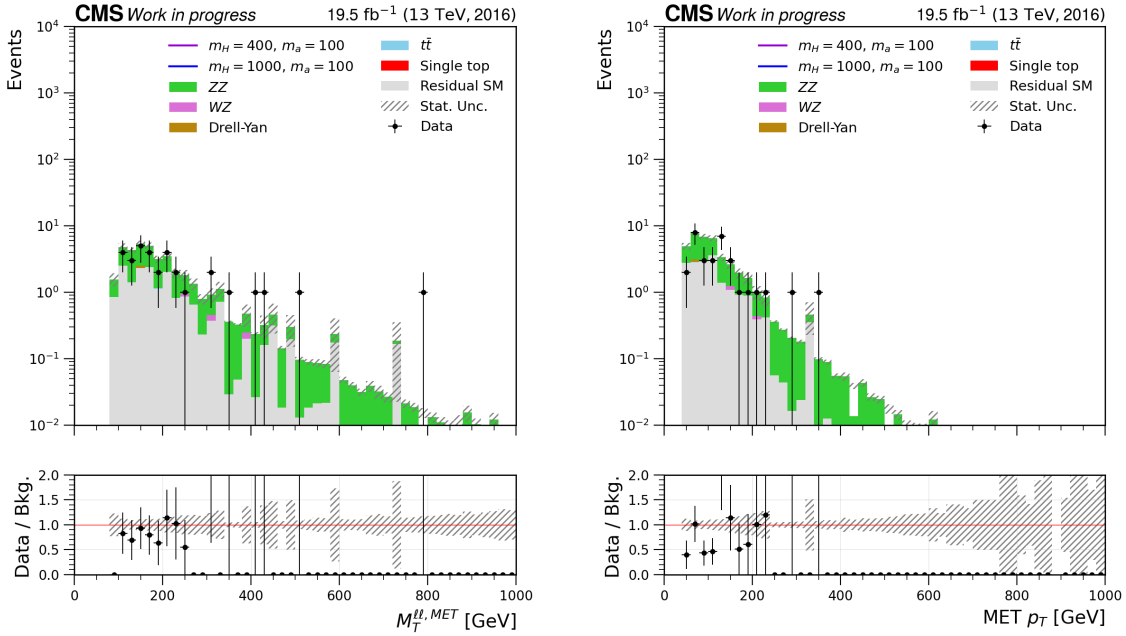
Fonte: O autor, 2023.

APÊNDICE H – Região de controle do ZZ nos demais períodos

Figura 47 - ΔR do sistema de dois léptons, p_T do sistema de dois léptons, massa transversa do sistema de dois léptons e $\not{E}_T$ e $\not{E}_T$ na região de controle do ZZ para o período de 2016 pre-VFP



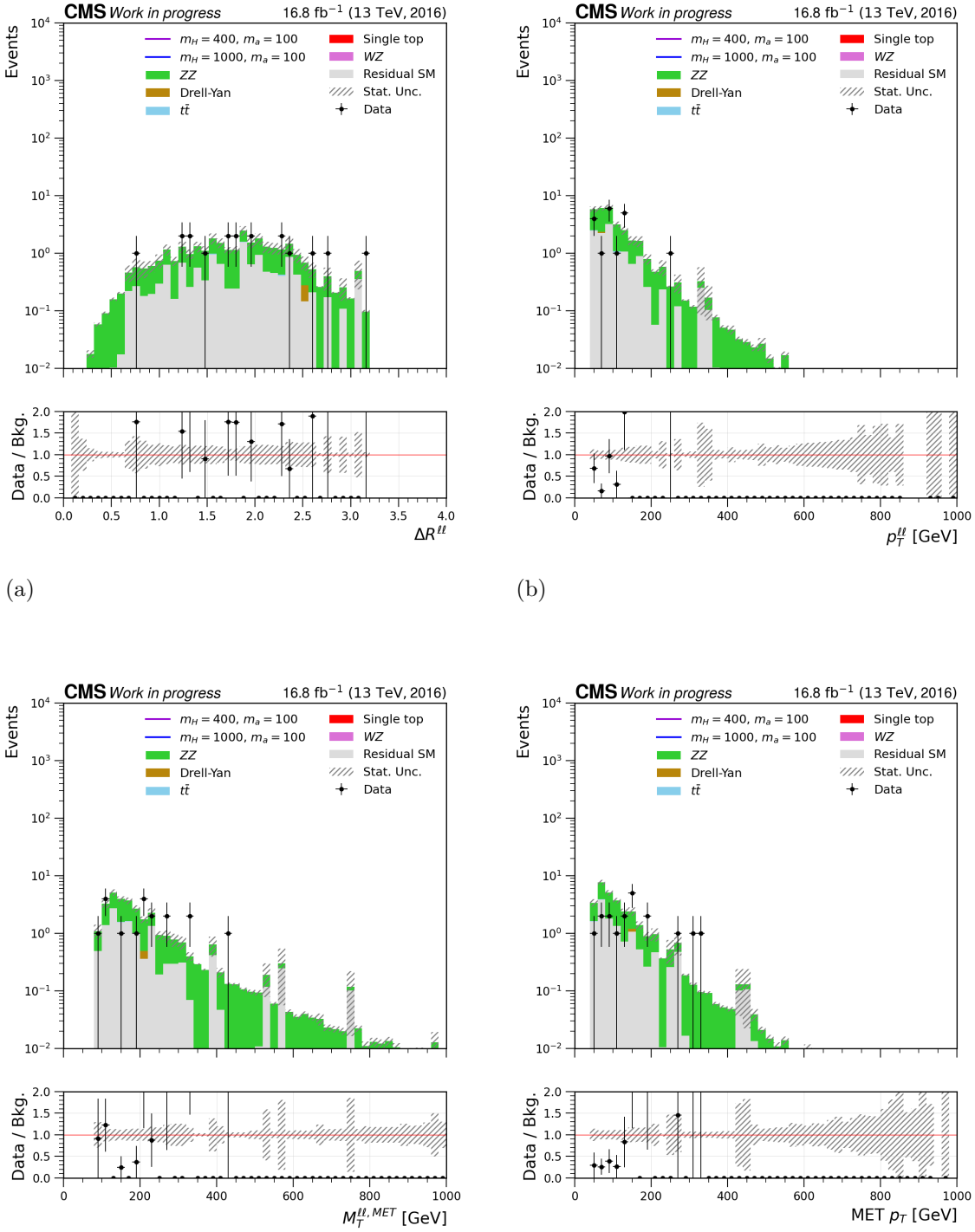
(a) (b)



(c) (d)

Fonte: O autor, 2023.

Figura 48 - ΔR do sistema de dois léptons, p_T do sistema de dois léptons, massa transversa do sistema de dois léptons e $\not{E}_T$ e $\not{E}_T$ na região de controle do ZZ para o período de 2016 post-VFP

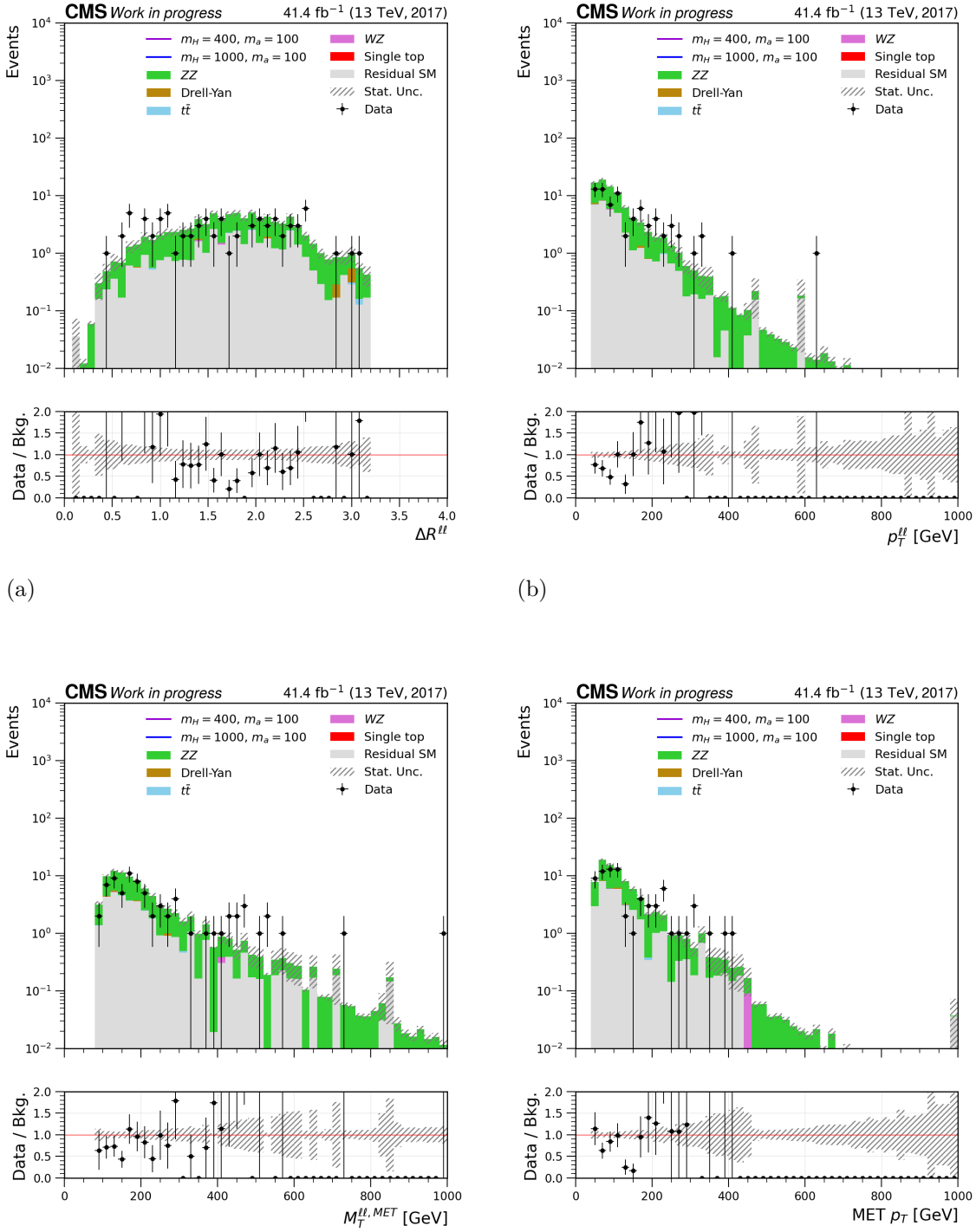


(a) (b)

(c) (d)

Fonte: O autor, 2023.

Figura 49 - ΔR do sistema de dois léptons, p_T do sistema de dois léptons, massa transversa do sistema de dois léptons e $\not{E}_T$ e $\not{E}_T$ na região de controle do ZZ para o período de 2017



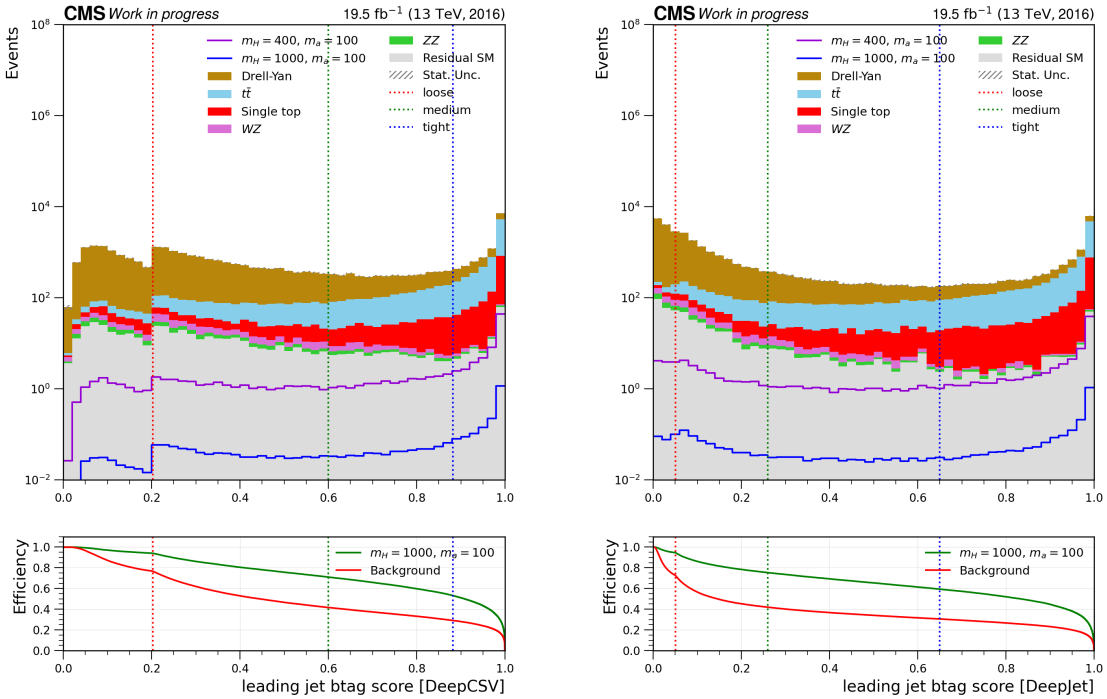
(a) (b)

(c) (d)

Fonte: O autor, 2023.

APÊNDICE I – Algoritmos de b-tagging nos demais períodos

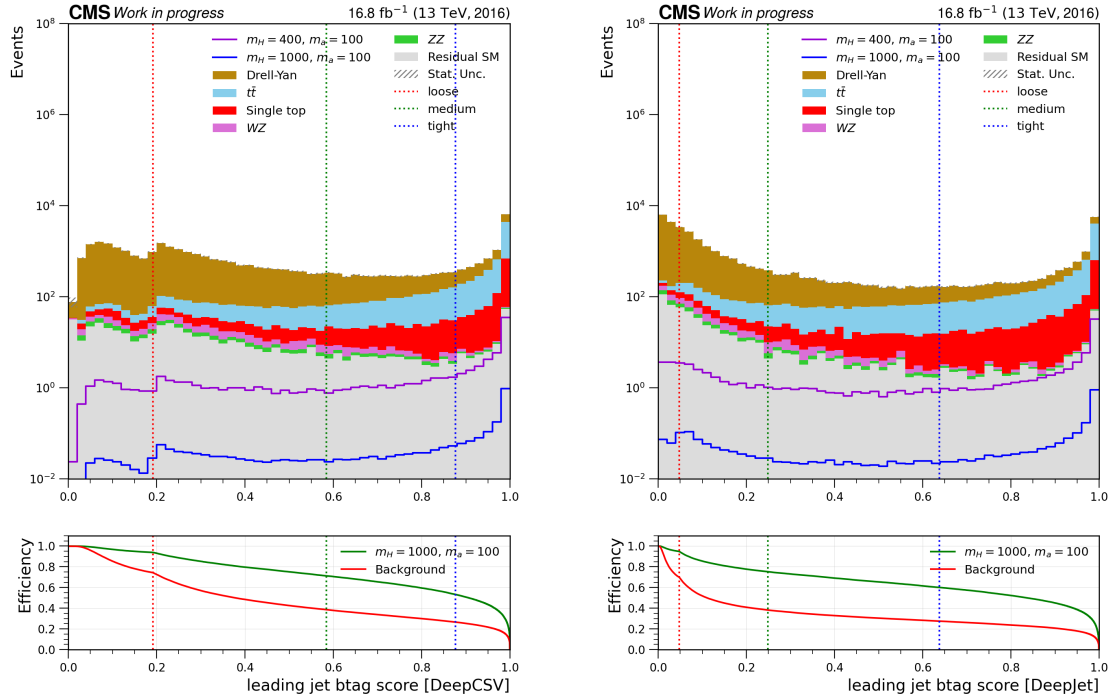
Figura 50 - Comparação dos algoritmos DeepCSV e DeepJet na região de sinal para o período de 2016 pre-VFP



(a) (b)

Fonte: O autor, 2023.

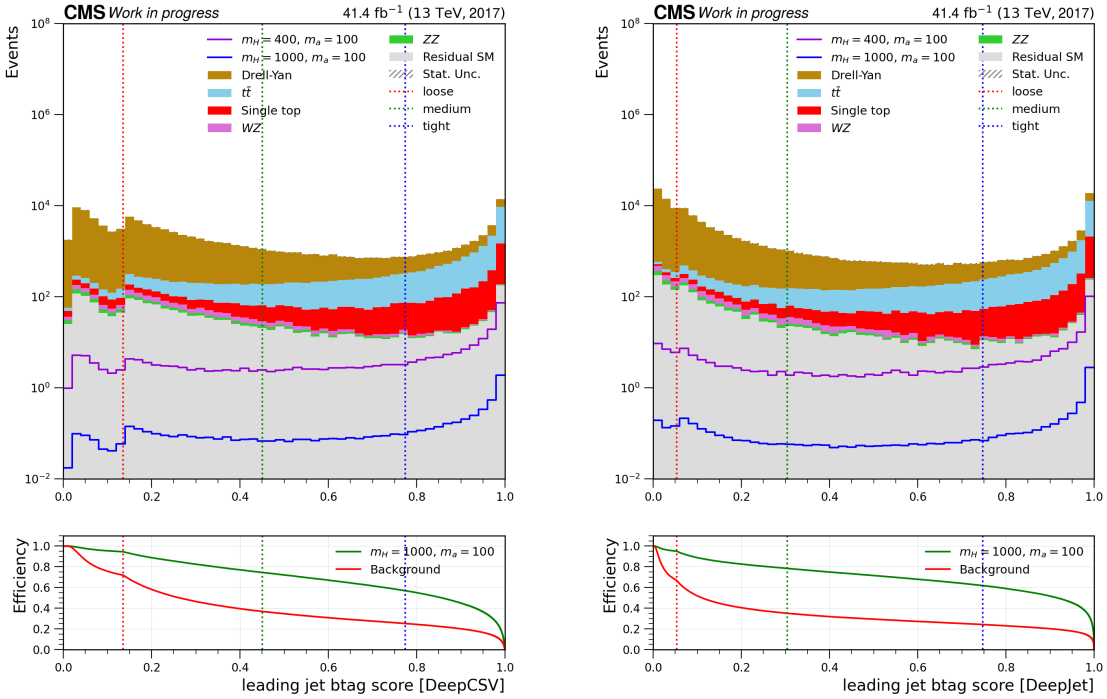
Figura 51 - Comparação dos algoritmos DeepCSV e DeepJet na região de sinal para o período de 2016 post-VFP



(a) (b)

Fonte: O autor, 2023.

Figura 52 - Comparação dos algoritmos DeepCSV e DeepJet na região de sinal para o período de 2017

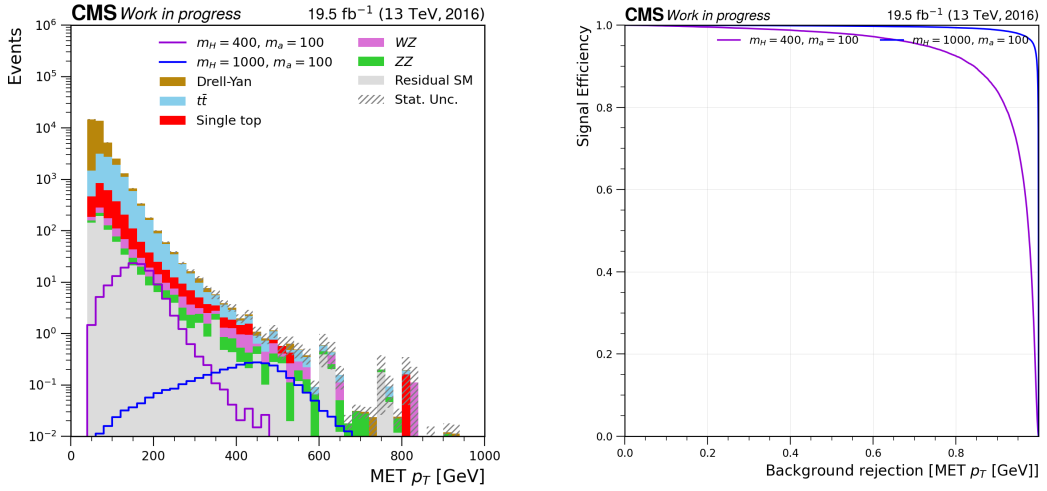


(a) (b)

Fonte: O autor, 2023.

APÊNDICE J – Seleção de *features* nos demais períodos

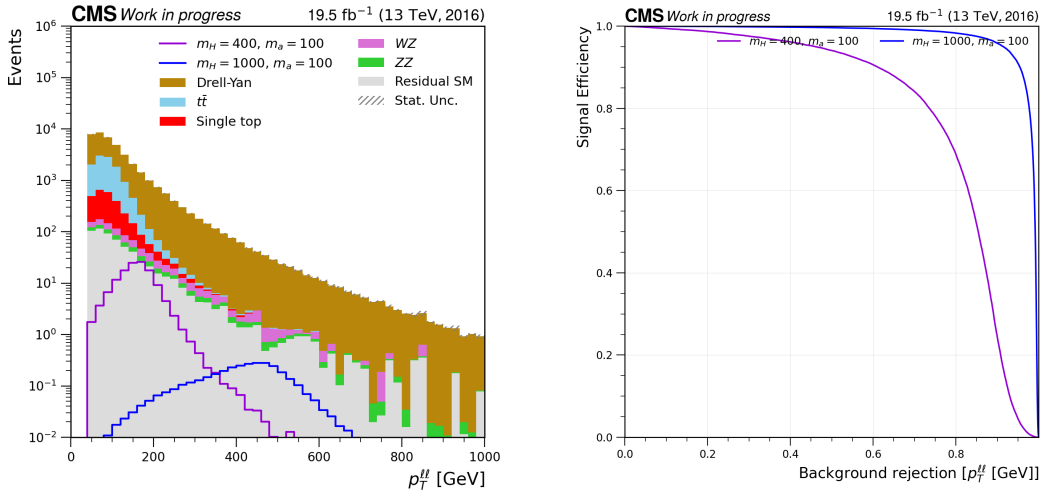
Figura 53 - Energia transversa perdida na região de sinal para o período de 2016 pre-VFP



(a) (b)

Fonte: O autor, 2023.

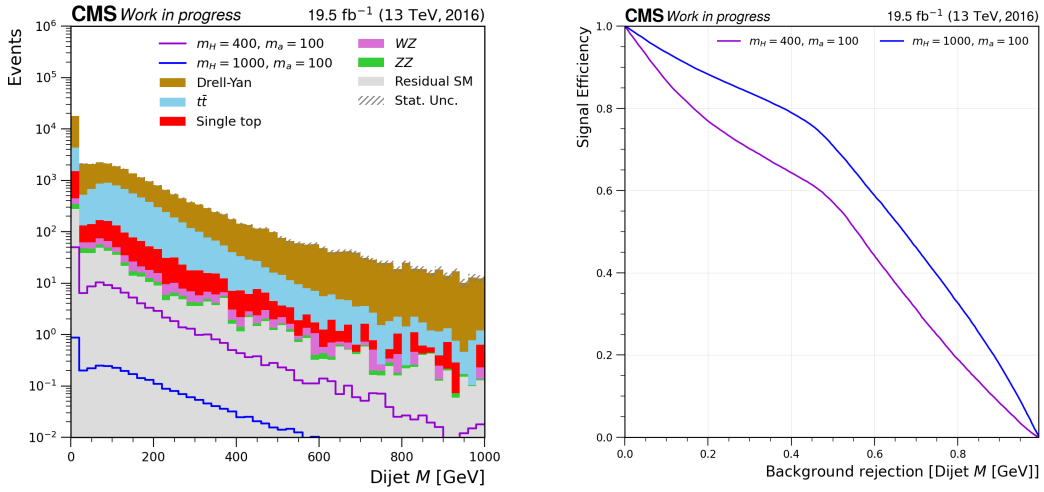
Figura 54 - Momentum transverso do sistema de dois léptons na região de sinal para o período de 2016 pre-VFP



(a) (b)

Fonte: O autor, 2023.

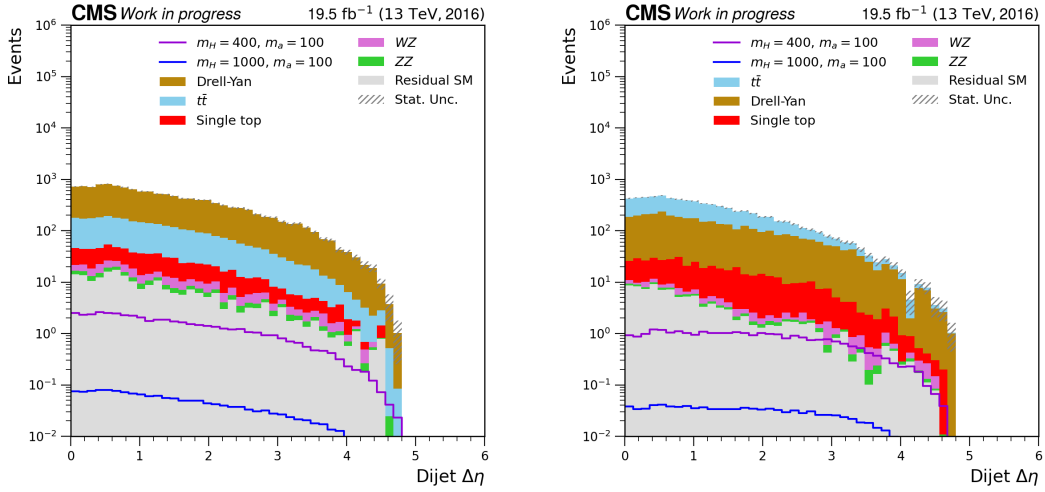
Figura 55 - Massa reconstruída do sistema de dois jatos na região de sinal para o período de 2016 pre-VFP



(a) (b)

Fonte: O autor, 2023.

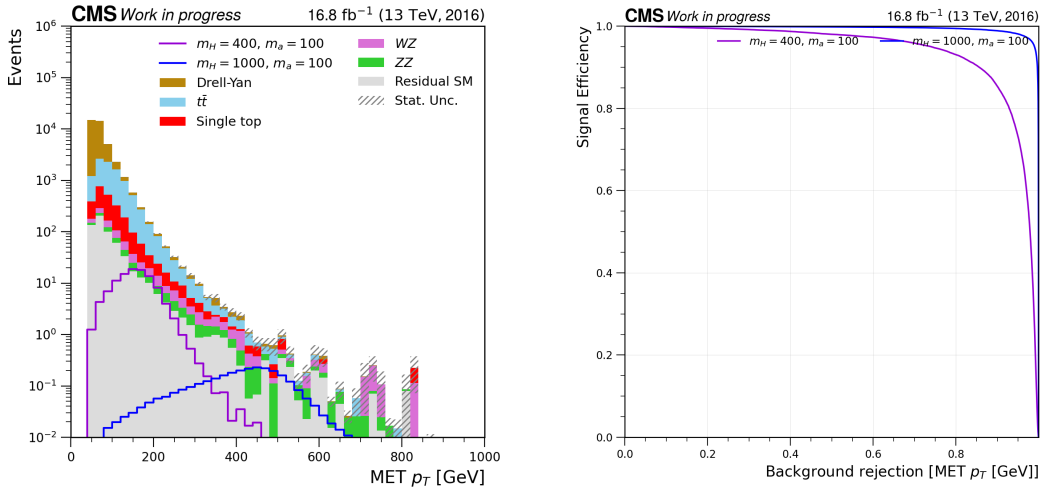
Figura 56 - Distribuição do $\Delta\eta$ do sistema de dois jatos na região de sinal para o período de 2016 pre-VFP



(a) (b)

Fonte: O autor, 2023.

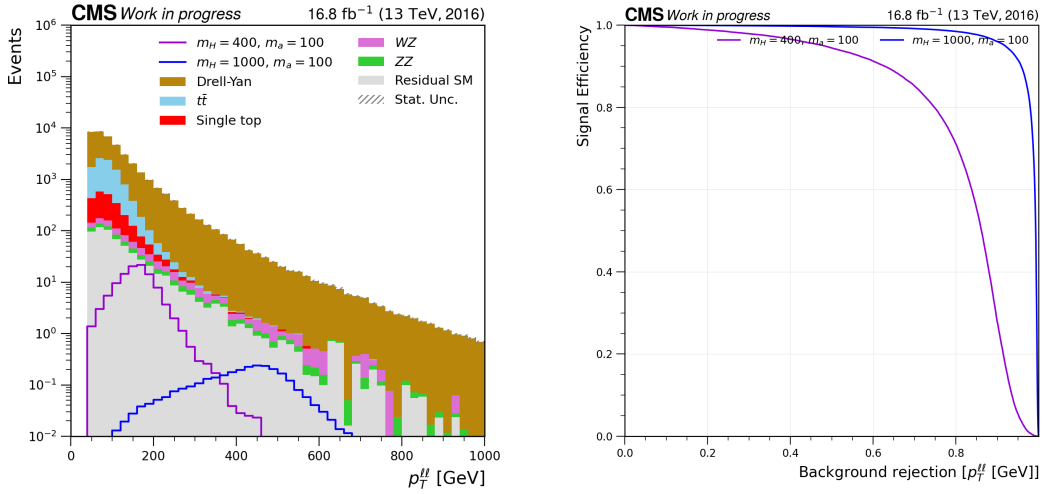
Figura 57 - Energia transversa perdida na região de sinal para o período de 2016 post-VFP



(a) (b)

Fonte: O autor, 2023.

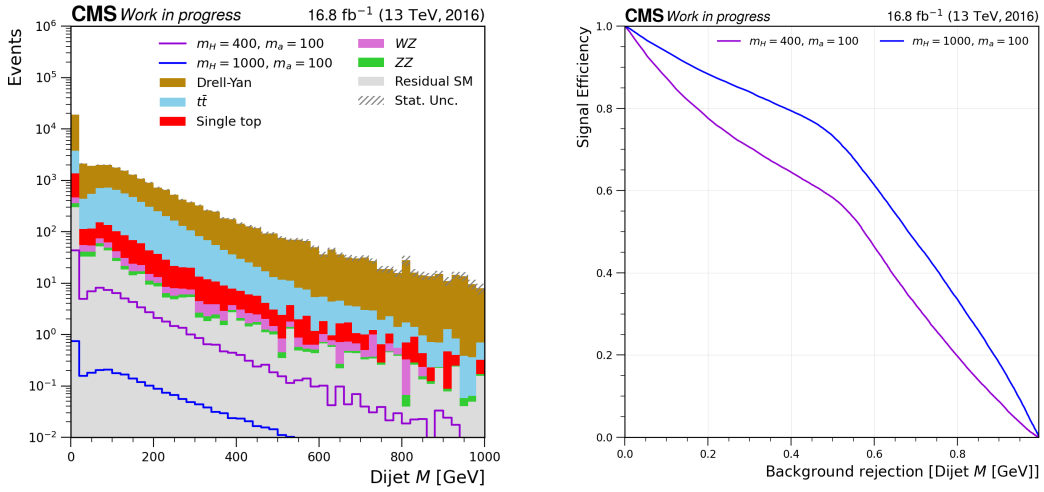
Figura 58 - Momentum transverso do sistema de dois léptons na região de sinal para o período de 2016 post-VFP



(a) (b)

Fonte: O autor, 2023.

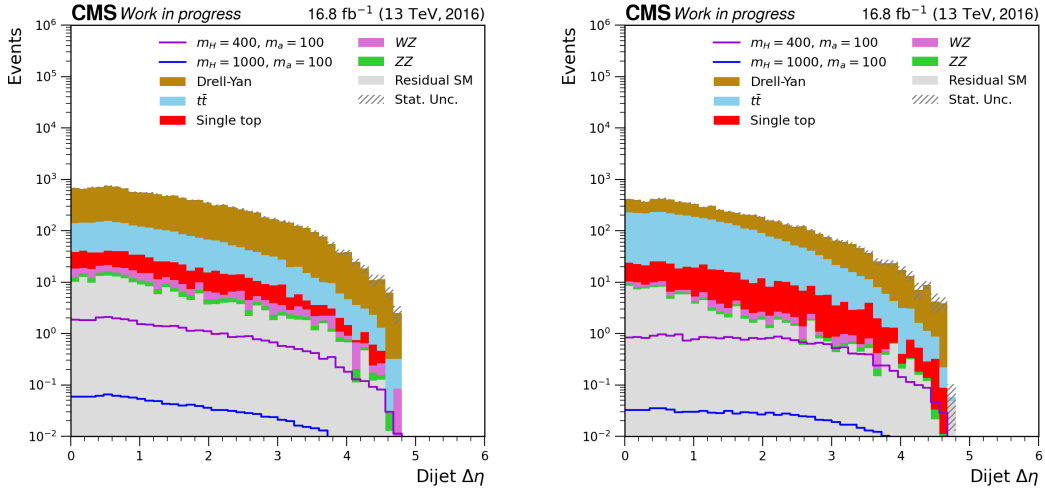
Figura 59 - Massa reconstruída do sistema de dois jatos na região de sinal para o período de 2016 post-VFP



(a) (b)

Fonte: O autor, 2023.

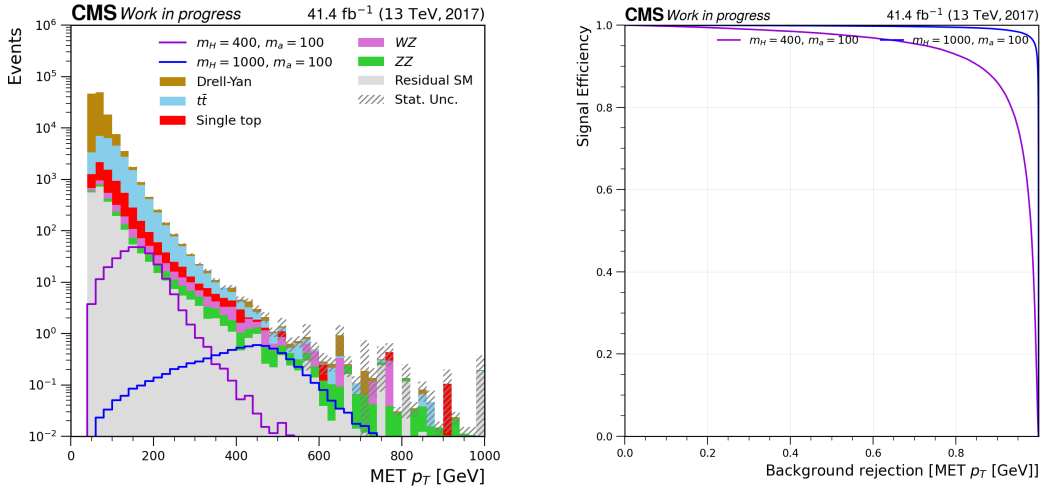
Figura 60 - Distribuição do $\Delta\eta$ do sistema de dois jatos na região de sinal para o período de 2016 post-VFP



(a) (b)

Fonte: O autor, 2023.

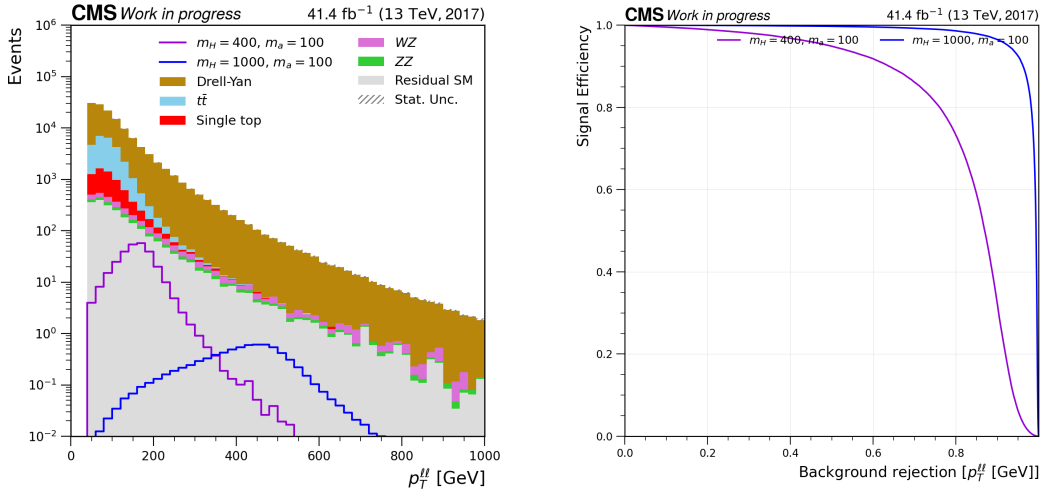
Figura 61 - Energia transversa perdida na região de sinal para o período de 2017



(a) (b)

Fonte: O autor, 2023.

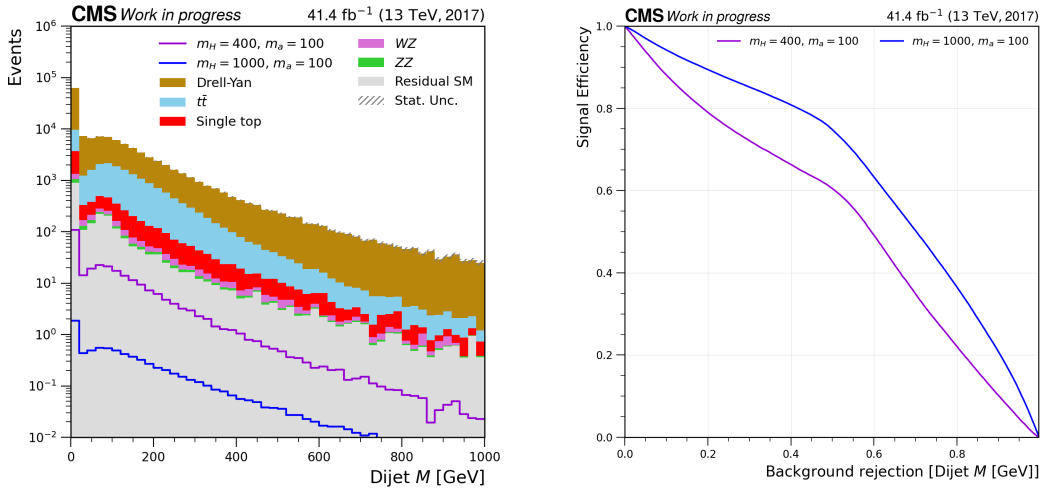
Figura 62 - Momentum transverso do sistema de dois léptons na região de sinal para o período de 2017



(a) (b)

Fonte: O autor, 2023.

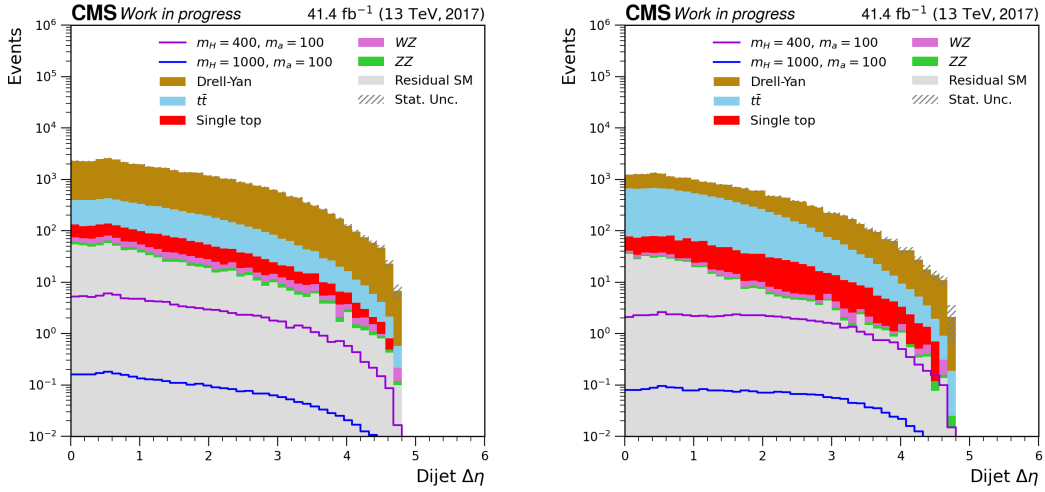
Figura 63 - Massa reconstruída do sistema de dois jatos na região de sinal para o período de 2017



(a) (b)

Fonte: O autor, 2023.

Figura 64 - Distribuição do $\Delta\eta$ do sistema de dois jatos na região de sinal para o período de 2017



(a) (b)

Fonte: O autor, 2023.

APÊNDICE K – *Features* utilizadas nos modelos de aprendizado de máquina

A lista a seguir enumera as *features* utilizadas para o treinamento de todos os modelos de aprendizado de máquina apresentados nesse trabalho para todos os períodos.

1. leading lepton p_T (Lépton com maior momentum transverso do evento)
2. $p_T^{\ell\ell}$
3. $|M_{\ell\ell} - M_Z|$
4. $\Delta R^{\ell\ell}$
5. $\not{E}_T$
6. $M_T^{\ell\ell,\not{E}_T}$
7. $\Delta\phi^{\ell\ell,\not{E}_T}$
8. trailing lepton p_T (Lépton com segundo maior momentum transverso do evento)
9. MT2LL (77)
10. Número de jatos provenientes do quark bottom

APÊNDICE L – Serviço no AlCaDB

Durante o Mestrado realizei uma breve participação no grupo AlCaDB do CMS, durante o período de preparativos para tomada de dados de 2022 (Run-3). Objetivo do trabalho era atualizar um script de validação de *triggers* no repositório AlCaTools (78) que naquele momento estava depreciado e não executava na última versão do CMSSW utilizado para o 2021 *Pilot Beam Test*. Uma série de modificações foram realizas no script para melhorar o *code design*, o funcionando com a versão mais recente do CMSSW e também a produção de resultados com maior interpretabilidade. O resultado desse trabalho foi apresentado em reuniões do AlCaDB e foi de extrema importância para validação rápida e confiável dos *triggers* em relação aos algoritmos de calibração dos dados.