

Universidade de São Paulo
Instituto de Física

Estudo da reação de breakup ${}^4\text{He}({}^{17}\text{F}, {}^{16}\text{O}+\text{p}){}^4\text{He}$
usando o alvo ativo pAT-TPC: uma abordagem
usando técnicas de Machine Learning

Guilherme Ferrari Fortino

Orientador: Prof. Dr. Valdir Guimarães _____

Coorientador: Dr. Juan Carlos Zamora Cardona _____

Dissertação de mestrado apresentada ao Instituto de
Física da Universidade de São Paulo, como requisito
parcial para a obtenção do título de Mestre(a) em Ciências.

Banca Examinadora:

Prof(a). Dr(a). Nome do(a) Professor(a) - Orientador (instituição de trabalho)

Prof(a). Dr(a). Nome do(a) Professor(a) (instituição de trabalho)

Prof(a). Dr(a). Nome do(a) Professor(a) (instituição de trabalho)

São Paulo
2022

Sumário

1	Introdução	9
2	O experimento	10
2.1	Produção do feixe secundário ^{17}F usando o sistema TWINSOL	10
2.2	O alvo ativo pAT-TPC	13
3	Desenvolvimento de ferramentas de <i>machine learning</i> para análise de dados	18
3.1	Tipos de redes neurais	19
3.2	Estrutura da rede neural	21
3.3	Sistemas de <i>machine learning</i>	23
3.4	Aplicações de <i>machine learning</i> na física nuclear	25
3.4.1	Análise de espectros para identificação de partículas (<i>particle identification</i> , PID)	25
3.4.2	Estimativa de raios e massas nucleares	26
3.4.3	Decaimento β e processo r	28
3.4.4	Alvos ativos	29
4	Reconstrução de nuvens de pontos a partir de algoritmos de <i>machine learning</i>	31
4.1	Construção do banco de dados para as redes neurais	32
4.1.1	Estimativa do fundo	33
4.1.2	Deconvolução do sinal	36
4.2	Análise dos pulsos com <i>machine learning</i>	38
4.2.1	Rede neural para o fundo	38
4.2.2	Rede neural para a deconvolução	41
4.2.3	Detecção de picos	44
4.2.4	Acoplando as redes neurais	47
5	Análise das nuvens de pontos	49
5.1	Detecção de trajetórias	49

5.2	Abordagens alternativas	55
5.2.1	Detecção de eventos com <i>machine learning</i>	55
5.2.2	Detecção de <i>outliers</i>	58
6	Resultados	60
7	Conclusão	62

Listas de figuras

2.1	Sistema TWINSOL à esquerda e pAT-TPC à direita. O feixe estável de ^{16}O entra à esquerda do TWINSOL, para então ser produzido o feixe secundário ^{17}F que irá ser conduzido até o alvo ativo pAT-TPC. Todo o sistema está localizado na University of Notre Dame.	11
2.2	Simulação computacional do sistema RIBRAS, onde as partículas carregadas em laranja surgem do ponto à direita daq figura e passam pelos dois solenoides em verde, para então serem focalizada em um ponto do plano focal em vermelho. A parte em preto dos solenoides corresponde aos limites físicos da bobina[5].	12
2.3	Valor do campo magnético B em tesla em função da posição em z em centímetro da bobina. A linha vertical tracejada preta indica o limite físico da bobina. O campo foi calculado à uma distância de 8 cm do eixo do solenoide. É possível ver claramente o efeito de borda que há em um solenoide finito[9].	12
2.4	Espectro biparamétrico $\Delta E - E$ de identificação de partículas. Nele é possível identificar que o feixe possui a presença de ^{17}F , ^{16}O e uma pequena parte de ^{17}O	13
2.5	Figura esquemática do pAT-TPC e o detector <i>Micromegas</i> [1].	14
2.6	Plano do <i>Micromegas</i> com a esquematização das <i>thick gems</i> . Os elétrons quando passam para um campo elétrico mais intenso ionizam o gás, produzindo ainda mais elétrons (evento chamado de avalanche de elétrons). Cada canal da eletrônica de saída produz um pulso como mostrado na parte de baixo da figura.	16
2.7	Evento reconstruído a partir da análise dos pulsos gerados pelo <i>Micromegas</i> . A cor representa a carga integrada de cada ponto de interação com o gás. . .	17

3.1	Exemplo de FFNN. A camada de entrada na esquerda propaga a informação para a direita (camada de saída). Todos os neurônios entre camadas estão conectados entre si.	19
3.2	Processo de convolução entre sinal azul em cima e o filtro em verde, resultando no sinal azul embaixo. A multiplicação é feita ponto a ponto e está indicada na caixa azul-clara.	20
3.3	Processo de convolução entre o sinal azul em cima e o filtro em verde, resultando no sinal azul embaixo. Agora são acrescentados zeros no inicio e no final do vetor para que o vetor saída tenha o mesmo tamanho do vetor de entrada (nesse caso 9).	21
3.4	Funções de ativação e seus respectivos gráficos.	22
3.5	Espectros biparamétricos $\Delta E \times E$, em 3.5a o espectro cru com uma pré marcação das partículas, e em 3.5b os conjuntos identificados usando <i>machine learning</i>	26
3.6	Histograma de identificação de partículas, a partir do conjunto 1 identificado na figura 3.5b. Em a temos o primeiro pico que corresponde ao estados fundamental do ^{14}N e em b temos o primeiro estado excitado do ^{14}N	26
3.7	Painel acima mostra a localização dos núcleos usados para o treino da rede neural. No painel de baixo temos o erro entre a previsão e o valor experimental da energia de ligação do núcleo para os núcleos usados como dados de validação. σ é o erro quadrático médio da rede neural.	27
3.8	Previsões para o raio de carga para isótopos do chumbo ($Z = 82$) para diferentes modelos teóricos ou que usam redes neurais. A previsão que contém barras de erro é a que foi brevemente descrita no texto.	28
3.9	Meias-vidas de decaimento β para isótonos com $N = 126$. A região hachurada verde mostra as previsões de uma rede neural. A região hachurada em azul mostra os resultados da mesma rede neural, porém seus dados de aprendizado são estendidos para incluir três meias-vidas extras de decaimento β para cada isótopo (indicado por círculos abertos) em direção à <i>drip-line</i> de nêutrons[44].	29

3.10 Projeções no plano xy de partículas dentro do alvo ativo, onde quanto mais escura for a cor, mais carga tem o ponto, e quanto mais clara a cor, menos carga tem o ponto. O objetivo da rede neural pode ser classificar corretamente se a imagem à direita corresponde à uma trajetória de um próton, carbono ou outra partícula, ou pode ser uma rede neural para classificação binária caso seja necessário classificar apenas entre próton ou carbono[45].	30
4.1 Exemplos de sinais produzidos pelos canais do detector. Em 4.1a o sinal possui apenas um pulso, enquanto em 4.1b há vários pulsos em sobreposição, formando um único pulso com largura maior que em 4.1a.	32
4.2 Ilustração que mostra a variação no formato da carga coletada a partir da passagem de uma partícula carregada dentro do TPC, onde o plano do detector está embaixo. No lado esquerdo de cada imagem, a distribuição do sinal coletado por um único pad (escuro) do plano de coleta é mostrado (o canal eletrônico de leitura é representado pela seta cinza em negrito). No caso de uma trajetória quase horizontal em relação ao plano do detector (a), o sinal é uma distribuição estreita, enquanto para uma trajetória próxima a uma direção vertical (ou perpendicular) em relação ao detector (b), a distribuição deve ser muito mais ampla (vários pontos de interação da partícula com o gás devem ser extraídos desse sinal). A última imagem ilustra o caso em mais de uma trajetória de partículas contribui para o sinal[14].	33
4.3 Histogramas com as respectivas <i>baselines</i> (linhas tracejadas) estimadas pelo método da convolução. O espectro resultante (sem o fundo) está em verde.	35
4.4 Histogramas com as respectivas <i>baselines</i> (linhas tracejadas) calculadas pelo <i>TSpectrum</i>	36
4.5 Histogramas sem as <i>baselines</i> antes (em azul) e depois da deconvolução (em vermelho). Os picos (em verde) e o limiar (linha tracejada preta) de detecção também estão indicados.	37
4.6 Arquitetura da rede neural que faz a inferência do fundo. O vetor de entrada deve ter dimensionalidade 512 x 1. Todas as partes com convolução não possuem o parâmetro <i>bias</i>	39

4.7	Resultados do treino da rede neural dada pela figura 4.6. A rede atingiu seu melhor resultado a partir da <i>epoch</i> 20 aproximadamente, quando começa um platô no <i>loss</i>	40
4.8	Exemplos da rede neural dada pela figura 4.6 em comparação com a saída do <i>TSpectrum</i>	41
4.9	Arquitetura da rede neural que faz a inferência da deconvolução do espetro. O vetor de entrada deve ter dimensionalidade 512 x 1. Todas as partes com convolução não possuem o parâmetro <i>bias</i>	42
4.10	Resultados do treino da rede neural dada pela figura 4.6.	43
4.11	Exemplos de deconvolução da rede neural dada pela figura 4.6.	43
4.12	Sinal após a deconvolução que mostra o pico detectado mais os pontos adicionais que irão facilitar o trabalho da rede neural (evitar o desbalanço de classe). Foram acrescentados 2 pontos à esquerda e à direita.	44
4.13	Arquitetura da rede neural que faz o recorte das regiões com picos. O vetor de entrada deve ter dimensionalidade 512 x 1.	45
4.14	Resultados do treino da rede neural dada pela figura 4.13.	45
4.15	Exemplos de detecção de picos usando a rede neural, em comparação com a detecção feita pelo algoritmo presente no SciPy, mostrada na figura 4.13. Os centroides detectados pela rede neural estão em vermelho, e os centroides detectados pelo SciPy estão em verde. Em azul está o espetro sem fundo após a deconvolução, resultantes	46
4.16	Arquitetura da rede neural que faz a inferência da <i>baseline</i> , em seguida faz a deconvolução do espetro sem o fundo e por fim faz a segmentação do sinal. O resultado da segmentação e da deconvolução são concatenados na parte final da rede neural. O vetor de entrada deve ter dimensionalidade 512 x 1.	47
4.17	Exemplos de eventos reconstruídos através da análise dos sinais com <i>machine learning</i> . A seta vermelha indica o sentido do feixe.	48
5.1	Sequência de análise de um evento. Em 5.1a temos o evento que é recebido para ser analisado, em 5.1b temos o mesmo evento após o HDB-SCAN (antes da correção) e 5.1c mostra depois da correção. As cores das retas são arbitrárias e servem apenas para a diferenciação.	51
5.2	Evento em que não foi detectado o feixe, apenas a partícula espalhada. O triângulo azul é o local calculo do vértice de reação dado pela equação 5.6.	53

5.3	Arquitetura da PointNet. A rede de classificação tem o <i>input</i> com n ponto com 3 coordenadas, onde são aplicadas sequencias de transformação que são agregadas por uma camada de max pooling. O <i>output</i> é a classificação para k classes possíveis. A rede de segmentação é uma extensão da rede de classificação, classificando ponto a ponto a nuvem de pontos, em m classes possíveis. Mais detalhes sobre a arquitetura podem ser encontrados na Ref. [67]	56
5.4	Nuvem de pontos que possui 3 trajetórias para serem detectadas e a rede neural de classificação calculou que haviam 3 trajetórias, o que indica um resultado correto do algoritmo.	58
5.5	Detecção de outliers com a rede neural de segmentação. A rede neural foi capaz de detectar os <i>outliers</i> desse evento com 95% de acurácia.	59
6.1	Histograma de comprimento de <i>track</i> no eixo y e ângulo de espalhamento no eixo x. O histograma foi feito coletando eventos que possuíam duas trajetórias com o mesmo vértice de reação, indicando a detecção simultânea do ^{16}O e do próton.	60

TabelasLista de algoritmos

Capítulo 1

Introdução

Um dos principais objetivos do estudo em física nuclear é entender a estrutura do núcleo. Apesar do sucesso do modelo de camadas em explicar as estruturas de núcleos estáveis, os núcleos instáveis ou exóticos, ricos ou pobres em nêutrons, continuam sendo um grande desafio para nossa compreensão.

Os núcleos leves radioativos são de grande interesse para a astrofísica nuclear.

Experimentos para o estudo desses núcleos

Este trabalho abordou a análise do experimento feito com o *prototype Active Target - Time Projection Chamber* (pAT-TPC) [1] usando técnicas de *machine learning*. A dissertação está esquematizada da seguinte forma: no capítulo 2 está feita a descrição do experimento; no capítulo 3 está feita uma breve descrição do que é *machine learning* com exemplos de aplicações em física nuclear; O capítulo 5 foi dedicado para a análise dos sinais (pulsos) gerados no experimento; No capítulo 5 está descrição a análise das nuvens de pontos reconstruídas a partir dos sinais; No capítulo 6 foi feita a construção das distribuições angulares; Por fim no capítulo 7 está a conclusão do trabalho.

Capítulo 2

O experimento

Esse capítulo descreve sobre a parte experimental desse trabalho que foi desenvolvido na Universidade de Notre Dame em Outubro de 2019. Nesse experimento o feixe radioativo ^{17}F foi produzido e selecionado por rigidez magnética pelo sistema TWINSOL (TWIN SOLenids)[2]. O alvo ativo *prototype Active Target - Time Projection Chamber* (pAT-TPC) [1] foi usado como alvo e detector simultaneamente. Na continuação, está descrito como foi feito o sistema de produção do feixe secundário ^{17}F e a detecção das reações induzidas por esse feixe no alvo ativo.

2.1 Produção do feixe secundário ^{17}F usando o sistema TWINSOL

A produção do feixe radioativo ^{17}F foi feita a partir da reação do feixe estável de ^{16}O com um alvo de deutério gasoso. O feixe primário ^{16}O foi produzido e acelerado por um acelerador tipo Tandem van de Graff do ISNAP (*Institute for Structure and Nuclear Astrophysics*) localizado na Universidade de Notre Dame, Estados Unidos, que possui uma tensão terminal de até 10 MV[3, 4]. O feixe de ^{16}O foi conduzido até a câmara alvo de produção, preenchida com ^4He , no início do sistema TWINSOL, onde partículas emergentes de reações nucleares surgiram (^{17}F , ^{16}O ^{17}O). A figura 2.1 mostra o sistema TWINSOL acoplado com o pAT-TPC.

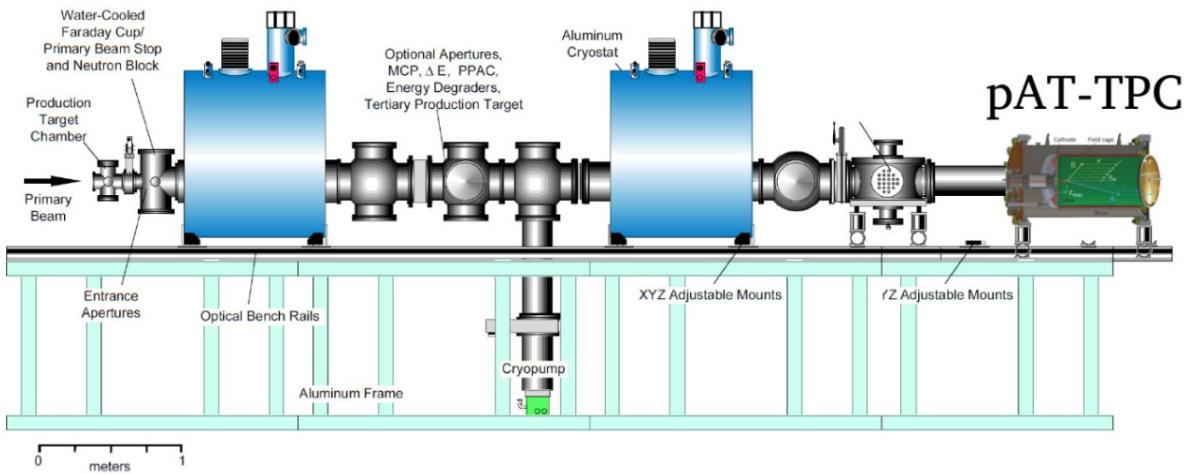


Figura 2.1: Sistema TWINSOL à esquerda e pAT-TPC à direita. O feixe estável de ^{16}O entra à esquerda do TWINSOL, para então ser produzido o feixe secundário ^{17}F que irá ser conduzido até o alvo ativo pAT-TPC. Todo o sistema está localizado na University of Notre Dame.

O TWINSOL é um sistema de produção de feixes radioativos em voo que possui dois solenoides supercondutores alinhados que são usados para produzir, coletar, transportar, focar e analisar feixes estáveis e radioativos. O sistema se baseia na seleção de partículas a partir da sua rigidez magnética ($B\rho$)[2, 5, 6]. Cada solenoide possui 30 cm de raio interno e 1 m de comprimento[2]. O fato de ser um solenoide finito faz com que surjam efeitos de borda na componente radial do campo magnético do solenoide, cujo efeito faz com que o solenoide seja capaz de focalizar partículas. Para entender melhor o efeito de borda no campo magnético, e consequentemente o funcionamento do TWINSOL, simulações computacionais usando a biblioteca GEANT4[7] foram feitas usando a geometria do sistema “irmão” do TWINSOL, o Radioactive Ion Beams in Brasil (RIBRAS), que também possui dois solenoides supercondutores alinhados[5, 8]. A figura 2.2 da geometria usada na simulação mostra os dois solenoides de cor verde focalizando as partículas de cor laranja em um ponto do plano focal em vermelho. O campo magnético usado na simulação é função da posição do eixo, de cada solenoide, e está mostrado na figura 2.3.

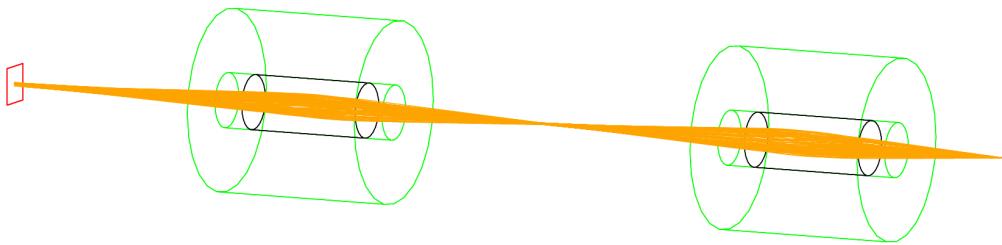


Figura 2.2: Simulação computacional do sistema RIBRAS, onde as partículas carregadas em laranja surgem do ponto à direita da figura e passam pelos dois solenoides em verde, para então serem focalizada em um ponto do plano focal em vermelho. A parte em preto dos solenoides corresponde aos limites físicos da bobina[5].

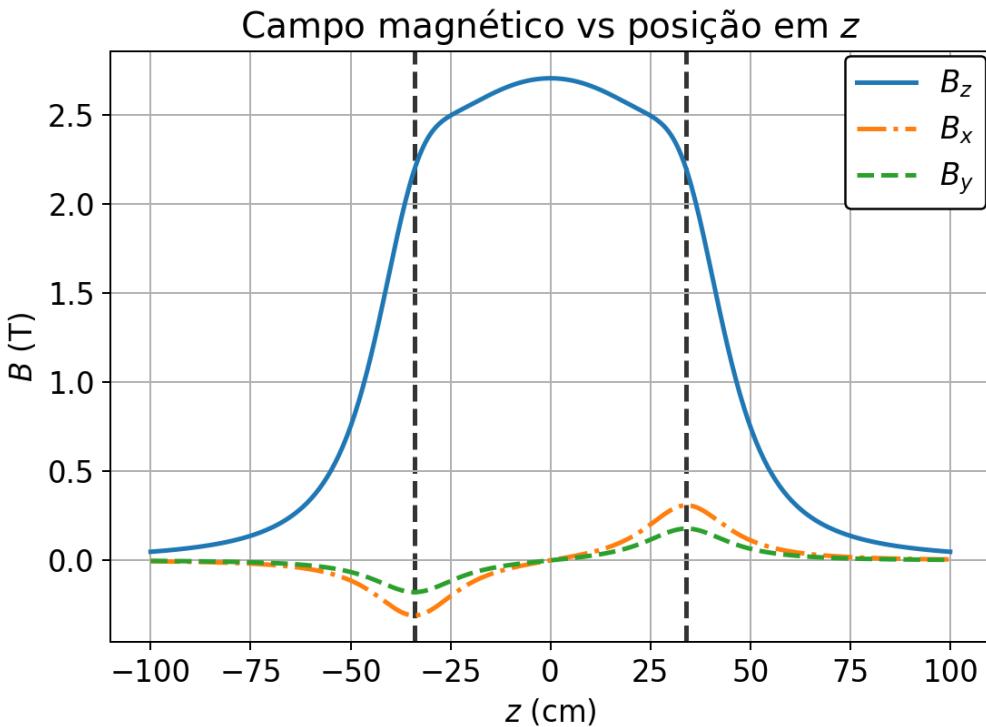


Figura 2.3: Valor do campo magnético B em tesla em função da posição em z em centímetro da bobina. A linha vertical tracejada preta indica o limite físico da bobina. O campo foi calculado à uma distância de 8 cm do eixo do solenoide. É possível ver claramente o efeito de borda que há em um solenoide finito[9].

A trajetória das partículas dentro do solenoide são helicoidais devido à força de Lorentz e possuem uma determinada frequência de ciclotrônico[5]. Além disso, cada solenoide é uma lente grossa usada para focalizar os feixes. Em uma aproximação de um solenoide como uma lente grossa, o foco depende da rigidez magnética da partícula através da relação[3, 6]:

$$\frac{1}{f} = \frac{B_z^2}{(B\rho)^2}, \quad (2.1)$$

onde f é o ponto focal, B_z a componente z do campo magnético, e $B\rho$ é dado por:

$$B\rho = \frac{mv}{q} = \frac{\sqrt{2mE}}{q}, \quad (2.2)$$

onde E é a energia, m sua massa e q seu estado de carga.

No experimento, os campos magnéticos dos solenoides foram ajustados para focalizar o feixe de ^{17}F dentro do pAT-TPC. No entanto, mesmo para partículas diferentes, o $B\rho$ pode ser muito próximo ou igual. Isso faz com que não seja possível obter um feixe de ^{17}F com 100% de pureza, e sim um coquetel de partículas[6]. O coquetel de partículas produzido possui 54% de ^{17}F , 41% de ^{16}O e cerca de 5% de ^{17}O . A figura 2.4 mostra o espectro biparamétrico de identificação de partículas, onde é possível identificar as partículas que estão presentes no feixe (coquetel de partículas). Por fim, o feixe produzido pelo TWINSOL é conduzido até o pAT-TPC.

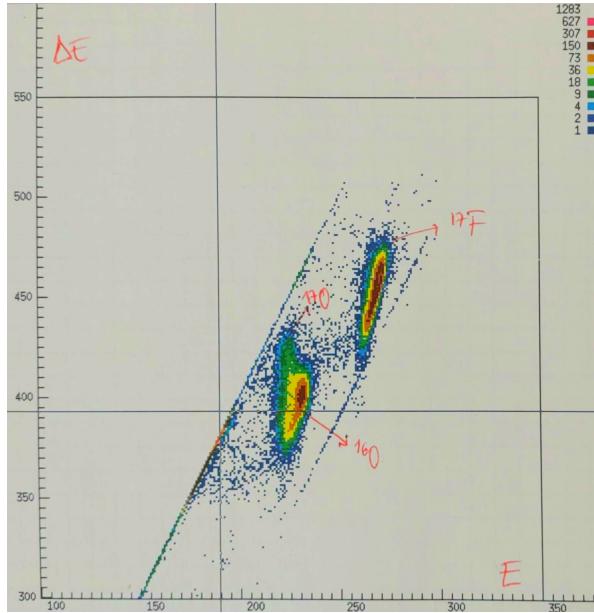
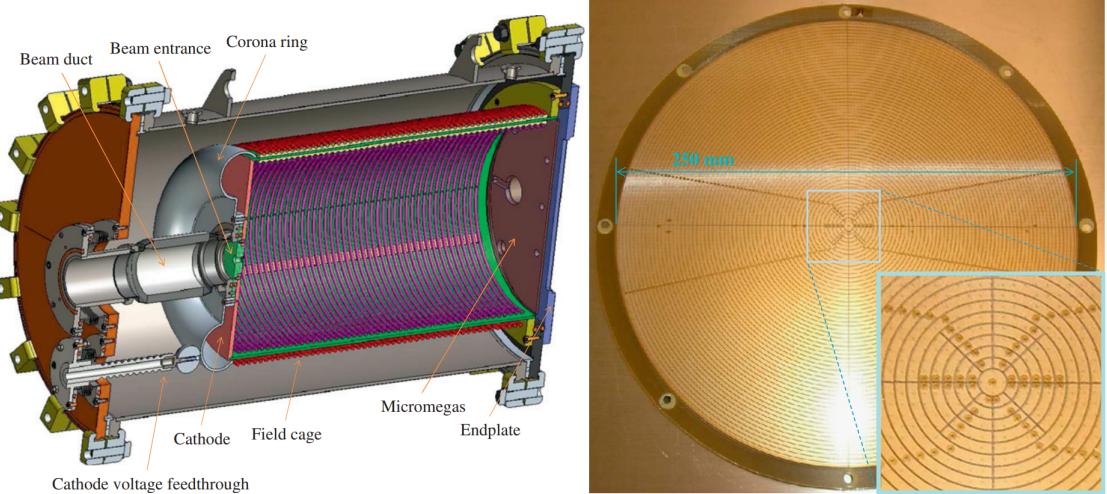


Figura 2.4: Espectro biparamétrico $\Delta E - E$ de identificação de partículas. Nele é possível identificar que o feixe possui a presença de ^{17}F , ^{16}O e uma pequena parte de ^{17}O .

2.2 O alvo ativo pAT-TPC

A figura 2.5a mostra o desenho esquemático do pAT-TPC. O detector possui uma cela cilíndrica de 50 cm de comprimento e 28 cm de diâmetro, onde o seu eixo é alinhado com o eixo do feixe[1], que entra pelo duto central. Nesse experimento, a câmara foi

preenchida com o ${}^4\text{He}$ gasoso puro à uma pressão de 350 Torr que serve tanto para alvo de reações nucleares, quanto para a própria medição e detecção dos produtos da reação[1, 10]. Tanto o feixe quanto partículas originadas da reação ionizam o gás e os elétrons que surgem dessa ionização foram conduzidos por um campo elétrico de 1 kV/cm perpendicular ao eixo da câmara até o plano detector (*pad plane*), o *Micromegas*[11], mostrado na figura 2.5b.



(a) Visão transversal do pAT-TPC. O gás é preenchido dentro da cela que possui um campo elétrico perpendicular ao plano do *Micromegas*, à direita da figura. O feixe incide na câmara entrando pelo duto de feixe à esquerda da figura.

(b) Foto do *Micromegas*. O detector é multi-pixelado com uma maior densidade no centro, parte destacada na imagem. O *pad* central tem diâmetro de 5 mm enquanto que as faixas coaxiais possuem passo de 2mm[12, 13].

Figura 2.5: Figura esquemática do pAT-TPC e o detector *Micromegas*[1].

O *Micromegas* é um dispositivo de amplificação de elétrons, que consiste em um plano detector com 2048 canais (*pads*) triangulares com eletrônica independente, que usa o *Generic Electronics for TPCs* (GET)[14]. Detalhes sobre a eletrônica podem ser encontrados nas Refs. [14, 13]. O formato triangular dos canais tem como objetivo maximizar a resolução espacial do detector[12]. Cada canal possui uma posição (x, y) fixa e a terceira coordenada z será determinada a partir do tempo de deriva dos elétrons no gás[1, 10, 12, 13]. Isso só é possível pois a velocidade de deriva (*drift*) dos elétrons é constante[15], portanto a posição em z da partícula é diretamente proporcional ao tempo de voo. Esse princípio que deu origem ao nome de *Time Projection Chamber*, pois o evento é projetado no tempo de deriva dos elétrons no gás. A equação 2.3 (equação de Langevin) descreve o movimento de um elétron com massa m e carga e é descrito por[15]

$$m \frac{d\vec{v}}{dt} = e \left(\vec{E} + \vec{v} \times \vec{B} \right) - \frac{m}{\tau} \vec{v}, \quad (2.3)$$

onde \vec{E} é o vetor campo elétrico, \vec{B} o vetor campo magnético, \vec{v} é o vetor de velocidade do elétron e τ é o tempo de colisão médio, que depende das propriedades termodinâmicas do gás. No caso deste experimento, \vec{B} é zero e a solução estacionária para a velocidade de *drift* do elétron é

$$\vec{v} = \frac{\tau}{m} e \vec{E}. \quad (2.4)$$

A velocidade de deriva depende das propriedades termodinâmicas do gás (temperatura, pressão) e também de sua condutividade elétrica[15]. Isso significa que a calibração da velocidade envolve não depender só do campo elétrico, mas também das propriedades do gás dentro do alvo ativo[1, 15]. Para acharmos a coordenada z , basta integrar a equação 2.4 para obter

$$z = \frac{\tau}{m} e \vec{E} (t - t_0), \quad (2.5)$$

onde no tempo $t_0 = 0$ o elétron está no plano do detector ($z = 0$).

O pAT-TPC conta com uma camada extra de *thick gems* acoplada ao detector micromegas. *Thick gems* usam do fato de que, no momento em que o elétron passa para uma região de campo elétrico ordens de grandeza maior que de sua origem, ocorre a ionização secundária (quando o elétron ioniza o gás). Isso provoca o que é chamado de avalanche de elétrons, amplificando a intensidade do sinal recebido[14]. A figura 2.6 mostra a esquematização do *Micromegas*, onde na eletrônica de saída é produzido um sinal em função do tempo.

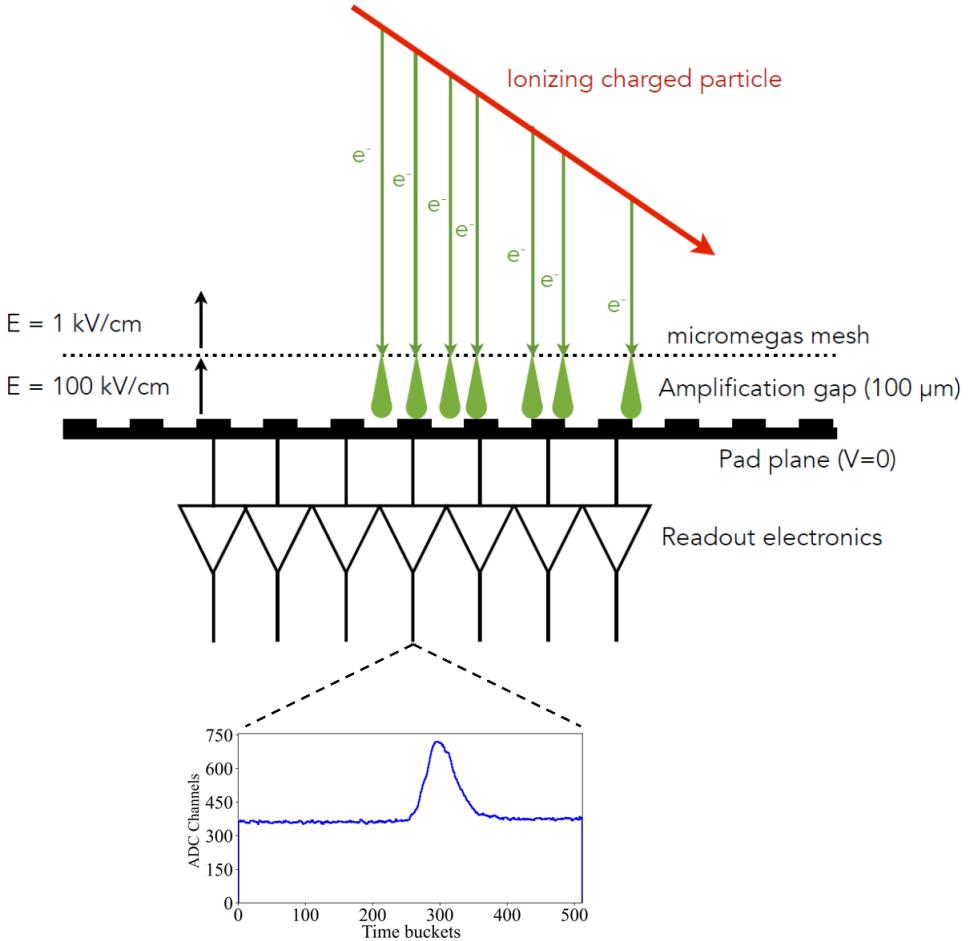


Figura 2.6: Plano do *Micromegas* com a esquematização das *thick gems*. Os elétrons quando passam para um campo elétrico mais intenso ionizam o gás, produzindo ainda mais elétrons (evento chamado de avalanche de elétrons). Cada canal da eletrônica de saída produz um pulso como mostrado na parte de baixo da figura.

O sinal é discretizado no tempo levando em conta a velocidade de deriva dos elétrons, dividindo em 512 canais o tempo que o elétron leva para percorrer toda câmara do TPC[13, 1]. A velocidade de deriva do elétron no gás ${}^4\text{He}$ é da ordem de $5 \text{ mm}/\mu\text{s}$ [1], onde o elétron percorre os 50 cm da câmara em cerca de $100 \mu\text{s}$. Dividindo esse tempo pelos 512 canais tem-se que cada canal (*time bucket*) possui 192 ns de largura. Cada centroide detectado é um ponto de interação de uma partícula carregada com o gás. A carga acumulada Q dessa interação é a área do pulso associado ao centroide. Cada centroide então representa um ponto no espaço (x, y, t, Q). Um exemplo de evento reconstruído está na figura 2.7.

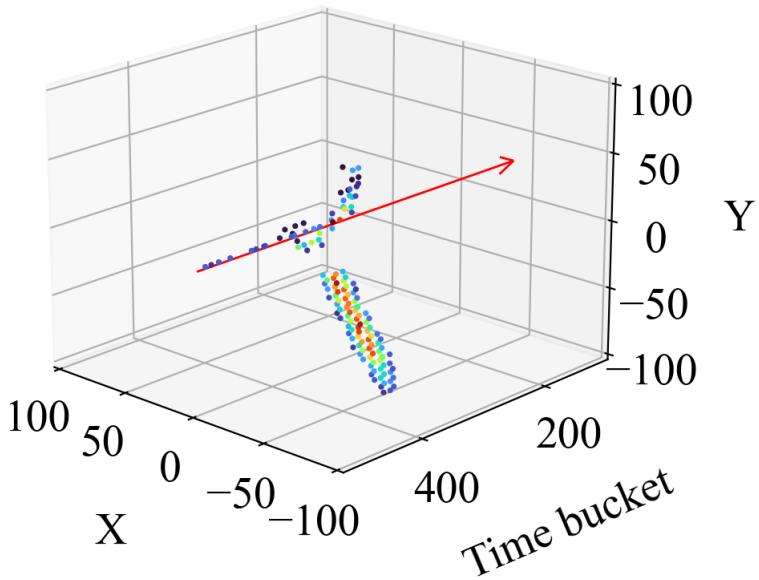


Figura 2.7: Evento reconstruído a partir da análise dos pulsos gerados pelo *Micromegas*. A cor representa a carga integrada de cada ponto de interação com o gás.

Para reconstituir eventos como o da figura 2.7, foram analisados cerca de 300 pulsos. Existem canais auxiliares que servem para evitar armazenar canais sem detecção. Caso haja detecção além do centro do *Micromegas* então os sinais gerados pelo evento são armazenados[13, 12]. O número de eventos reconstruídos é da ordem de milhões, portanto a quantidade de sinais que precisam ser analisados é muito grande, em comparação com experimentos em física nuclear com apenas alguns canais de detecção, o que gera a necessidade de desenvolvimento de algoritmos extremamente eficientes em tempo para que a análise não demande muito tempo. Para a análise completa do experimento foram seguidas as seguintes etapas:

- Análise dos pulsos de cada interação das partículas com o gás. Isso envolve remover o fundo, localizar os picos e obter os tempos e carga integrada de cada caso;
- Reconstruir eventos em 3D (nuvens de pontos) a partir da análise de sinais. As nuvens de pontos precisam ser analisadas com algoritmos de reconhecimento de padrões que permitem ajustar as trajetórias das partículas em 3D;
- Reconstruir a cinemática das partículas com as trajetórias e energia depositada no gás. Isto permite obter as distribuições angulares.

A análise dos pulsos, reconstituição de eventos, reconstrução da cinemática, gráficos das distribuições angulares e resultados são apresentadas nos próximos capítulos.

Capítulo 3

Desenvolvimento de ferramentas de *machine learning* para análise de dados

Um dos objetivos desse trabalho está em desenvolver ferramentas baseadas em *machine learning* para a análise do grande volume de dados gerada pelo alvo ativo. Nesse capítulo está explicada a metodologia usada para a implementação dessas ferramentas, e algumas aplicações na física nuclear.

Machine learning é a área de estudo que desenvolve algoritmos para que eles possam aprender com os dados, sem serem explicitamente programados para isso[16]. Supõe-se que uma rede neural imite um sistema biológico, em que os neurônios interajam enviando sinais na forma de funções matemáticas entre as camadas. Isso inspirou o uso do modelo matemático simples de uma função linear nos parâmetros para um neurônio artificial[17]:

$$y = f \left(\sum_{i=1}^n \omega_i x_i + b_i \right) = f(z), \quad (3.1)$$

onde y é a saída do neurônio, que corresponde à função de ativação f que depende da soma ponderada, onde o peso é ω_i , das entradas x_i dos outros n neurônios. O termo b_i corresponde ao parâmetro *bias*. A ideia é fazer um neurônio receber a informação de todos os outros neurônios da camada anterior, fazendo uma média ponderada (onde o peso que será estimado pelo algoritmo de *machine learning*) e somando com um termo independente (*bias*, que também é estimado). Os parâmetros ω_i e b_i serão estimados através de um determinado procedimento, chamado de minimização (o treino da rede neural).

3.1 Tipos de redes neurais

Uma rede neural artificial, *Artificial Neural Network* (ANN), é um modelo computacional que consiste de camadas de neurônios. ANNs foram desenvolvidas para o estudo de inteligência artificial[16, 18]. ANNs consistem principalmente numa camada de entrada (*input layer*), uma camada de saída (*output layer*) e eventuais camadas entre essas duas, chamadas de camadas ocultas (*hidden layers*). Os tipos mais comuns são:

Feed-Forward Neural Networks

A *Feed-forward neural networks* (FFNN) é a primeira e mais simples rede neural desenvolvida[19, 20]. Nessa rede a informação se move apenas para frente através de camadas (da camada de entrada até a camada de saída). A figura 3.1 mostra uma representação de rede, onde os neurônios são representados por círculos, enquanto que as linhas mostram as conexões entre os neurônios. Cada neurônio recebe informação de todos os neurônios da camada anterior, portanto a rede é chamada de totalmente conectada, *fully-connected* (FC), FFNN.

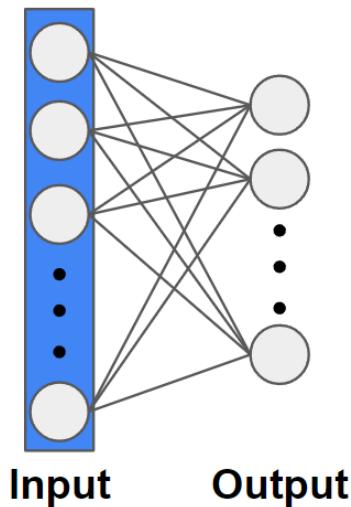


Figura 3.1: Exemplo de FFNN. A camada de entrada na esquerda propaga a informação para a direita (camada de saída). Todos os neurônios entre camadas estão conectados entre si.

Convolutional Neural Network

Uma variante da FFNN é a chamada de rede neural convolucional, *convolutional neural network* (CNN). Do ponto de vista matemático sobre convoluções, a convolução descrita como $(f * g)(t)$ de uma função $f(t)$ e outra $g(t)$ é definida como:

$$(f * g)(t) \equiv \int_{-\infty}^{\infty} f(\tau)g(t - \tau)d\tau. \quad (3.2)$$

Para o caso discreto, com g sendo uma função resposta finita de tamanho $2M$, temos

$$(f * g)[n] = \sum_{m=-M}^{M} f[n-m]g[m]. \quad (3.3)$$

Convoluçãoções são invariantes sobre rotação e translação, portanto são muito utilizadas para processamento de sinais e imagens[21]. Para a convolução discreta se escolhe um filtro que irá atuar no vetor desejado. Para ilustrar o que significa isso, no caso discreto e unidimensional, a figura 3.2 mostra o processo de convolução de um vetor de tamanho 9 e um filtro de tamanho 3.

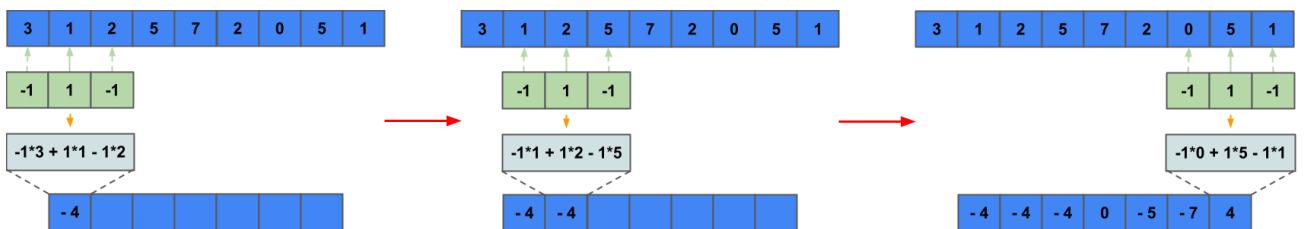


Figura 3.2: Processo de convolução entre sinal azul em cima e o filtro em verde, resultando no sinal azul embaixo. A multiplicação é feita ponto a ponto e está indicada na caixa azul-clara.

Percebe-se que o sinal resultante tem dimensão menor que o sinal original. O filtro (também chamado de *kernel*) atua em um ponto que possua vizinhos o suficiente para o restante do filtro poder fazer a multiplicação ponto a ponto. Esse tipo de convolução tem o chamado emparelhamento válido (*valid padding*). O tamanho n_2 resultante do vetor de saída é

$$n_2 = n_1 - m + 1, \quad (3.4)$$

onde n_1 é o tamanho do vetor de entrada e m o tamanho do filtro (*kernel size*). Para que o vetor de saída tenha o mesmo tamanho do vetor de entrada, são acrescentados zeros em torno da entrada, de forma que a saída tenha o mesmo tamanho da entrada. Esse é o chamado emparelhamento igual (*same padding*). A figura 3.3 ilustra esse processo.

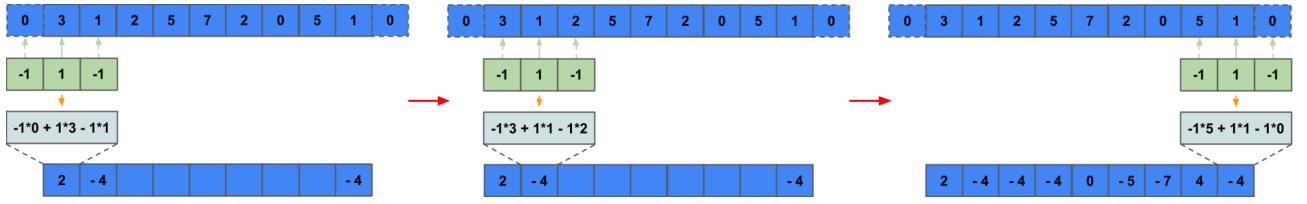


Figura 3.3: Processo de convolução entre o sinal azul em cima e o filtro em verde, resultando no sinal azul embaixo. Agora são acrescentados zeros no inicio e no final do vetor para que o vetor saída tenha o mesmo tamanho do vetor de entrada (nesse caso 9).

Como estamos no contexto de *machine learning* (inteligência artificial), *a priori* não sabemos quais os valores dos filtros que devem ser aplicados, apenas seus tamanhos e como agem. A ideia é estimar os valores do filtro (através do treino da rede neural) que deve ser aplicado para se obter o resultado desejado.

Cada filtro aplicado gera um mapa característico (*feature map*), que é o resultado da atuação do filtro em um vetor. Usualmente, em uma CNN se escolhe o tamanho do filtro, *padding* (*valid* ou *same*) e quantos filtros serão aplicados (para saber quantos *feature maps* serão gerados). Como temos vários mapas gerados por cada filtro, isso acarreta em um aumento de dimensionalidade. Para filtrar/selecionar os mapas é usado um critério, como por exemplo selecionar valores máximos dos mapas gerados dada uma janela de atuação (quantos mapas serão comparados para selecionar o máximo valor). O *Max-Pooling* faz isso, selecionando valores máximos para uma determinada quantidade de mapas sendo comparados (*pool size*).

Existem outros tipos de redes neurais que não estão discutidas aqui, porém podem ser encontradas nas referências [22, 23].

3.2 Estrutura da rede neural

Para a construção de uma rede neural (nesse caso em específico de uma rede neural supervisionada, que está discutida mais para frente), é preciso definir sua estrutura. A estrutura da rede neural possui camadas e cada camada pode possuir uma função de ativação. No geral, tanto o *input* quanto o *output* possuem dimensão fixa. A rede é treinada minimizando uma função custo, otimizando os parâmetros da rede através de um otimizador.

Para cada camada da arquitetura devemos escolher sua função de ativação. Tanto FNNNs quanto CNNs podem possuir funções de ativação (função f da equação 3.1). Dentre muitas funções de ativação podemos citar a *Rectified Linear Units* (ReLU)[24], sigmoide[25], linear e tangente hiperbólica[26]. A figura 3.4 mostra os gráficos dessas

funções de ativação, onde no eixo x é o argumento e no eixo y o resultado da função.

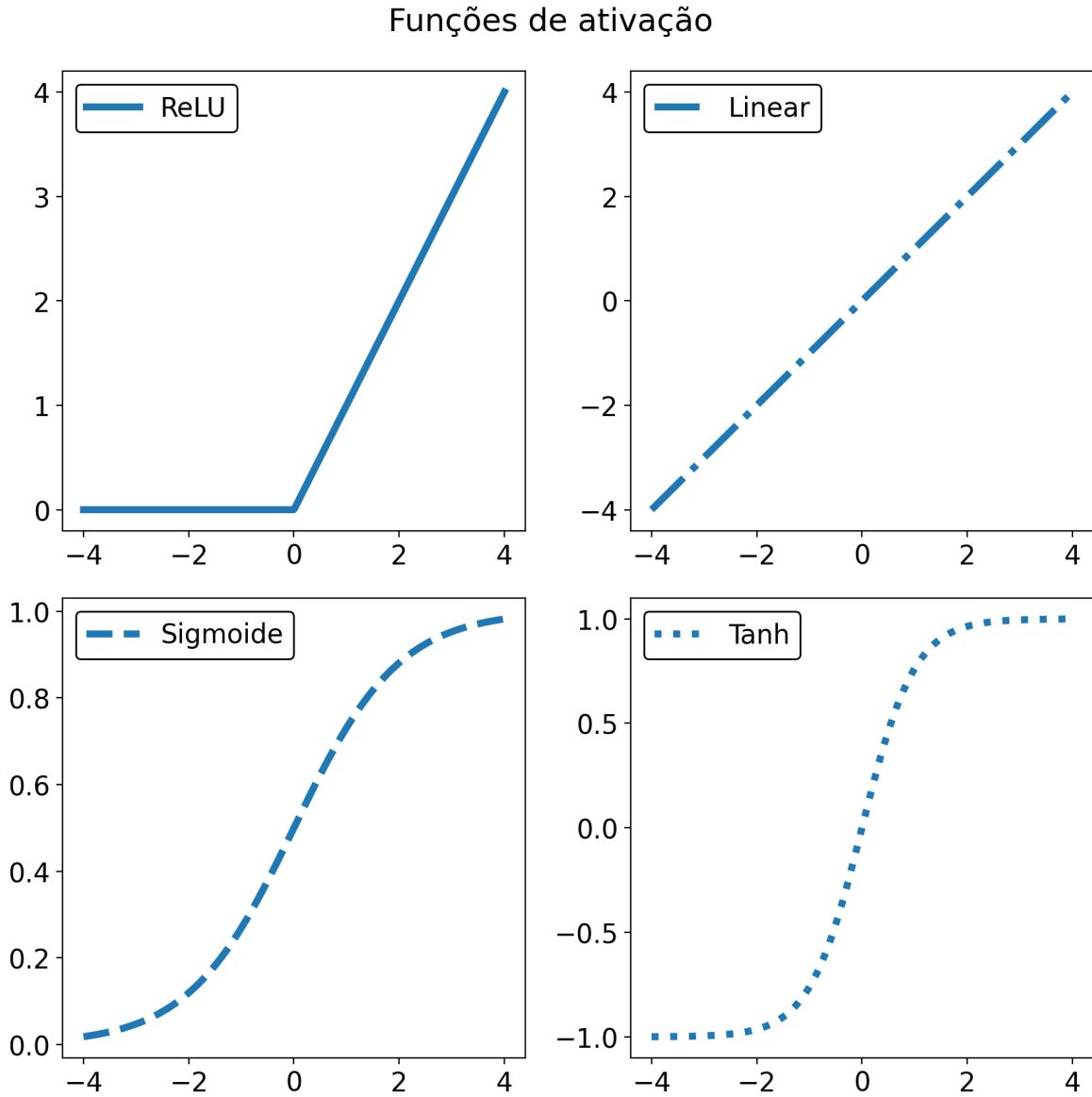


Figura 3.4: Funções de ativação e seus respectivos gráficos.

O próximo passo é definir a função custo (também chamada de *loss*) e o otimizador. A função custo tem o papel de retornar valores altos para previsões erradas e valores baixos para previsões corretas. Por exemplo, se queremos treinar uma rede neural para classificação binária (que prevê duas saídas possíveis), devemos usar a função custo chamada de *binary cross-entropy* dada por[27]

$$C(p(y_i)) = -\frac{1}{N} \sum_{i=1}^N y_i \log(p(y_i)) + (1 - y_i) \log(1 - p(y_i)), \quad (3.5)$$

onde y_i é o rótulo (*label*), $p(y_i)$ é a probabilidade do ponto y_i ser 1 e N é o número

de pontos. O objetivo da rede neural é achar o mínimo da função $C(p(y_i))$, o que implica diretamente na melhor solução para o conjunto de dados. Isso é feito pelo método de retropropagação do erro (*backpropagation*[28]) por um otimizador. Outros exemplos de *loss* são o erro quadrático médio ou *categorical cross-entropy*[29].

O otimizador tem o objetivo de otimizar os parâmetros presentes na rede neural, buscando o mínimo global da função custo, o que nem sempre acontece, pois a minimização pode parar em um mínimo local da função. Existem diversos otimizadores, como por exemplo o *Stochastic Gradient Descent* (SGD), ADAM, ADAMAX[30], entre outros[31]. Para o SGD, temos que a atualização de parâmetros é dada por

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} C(\theta), \quad (3.6)$$

onde θ_j é o parâmetro a ser atualizado, α é a *learning rate* e $C(\theta)$ é a *loss* que depende dos parâmetros θ .

Para enfim treinar a rede neural, se escolhe o *batch size*, que é o tamanho de amostras que será usada para o treino, por iteração em cada rodada de treino (*epoch*). Por exemplo, se usamos 1000 dados para o treino, e o *batch size* é 500, cada *epoch* terá duas iterações. No geral, se usam *batch sizes* pequenos, pois o consumo de memória é mais eficiente.

Para avaliação do modelo se usam dados de validação, que servem para verificar o comportamento da rede neural que está sendo treinada. Esse conjunto de dados usados não é usado para o treino, são usados apenas para verificar possíveis problemas como o *overfit*. O *overfit* ocorre quando a rede neural começa a se adequar perfeitamente aos dados de treino, perdendo a capacidade de previsão em dados que não estão sendo usados no treino pela rede neural.

Além dos dados de validação, podemos escolher métricas que auxiliam a visualização do comportamento da rede neural durante o treino e nos retornam informações importantes sobre sua qualidade. Exemplos importantes de métricas são: acurácia binária, erro médio absoluto e acurácia categórica[32]. Por exemplo, caso seja necessário verificar se uma rede neural está fazendo previsões certas em um problema cuja classificação é binária, então a métrica deve ser a acurácia binária. Tudo depende do objetivo da rede neural.

3.3 Sistemas de *machine learning*

Podemos dividir os sistemas de *machine learning* em quatro tipos:

Aprendizado Supervisionado

Aprendizado supervisionado é quando fornecemos para a rede neural um conjunto de dados para o treino com a solução desejada (chamados de *labels*). Um uso típico é para problemas de classificação[16]. Por exemplo, classificação de imagens (identificação de figuras), previsão de valores numéricos etc. Exemplos de algoritmos supervisionados são:

- *k-Nearest Neighbors*[33]
- Regressão linear
- *Support Vector Machines* (SVMs)
- *Decision Trees and Random Forests*
- Redes neurais

Aprendizado não supervisionado

Aprendizado não supervisionado é quando fornecemos o conjunto de dados para o treino, porém sem solução. A ideia é aprender sem supervisão. Um problema comum, por exemplo, é quando queremos identificar *clusters* em um conjunto de dados (*clustering*)[34, 35].

Aprendizado semi supervisionado

Aprendizado supervisionado é quando apenas parte do conjunto de dados para o treino possui *labels*. Isso é comum quando se obtém conjuntos de dados diferentes e apenas parte deles foi classificado[36].

Aprendizado por reforço

Aprendizado por reforço é quando um sistema, chamado de *agente* nesse contexto, aprende através do ambiente, realizando ações que maximizam sua recompensa. Por exemplo, caso o sistema realize uma ação incorreta, ele recebe uma penalidade, fazendo com que procure outra maneira de realizar a ação, dessa vez de maneira correta, para poder ganhar uma recompensa[37]. Esse tipo de sistema é muito usado, por exemplo, em automatização robótica, como carros que pilotam sozinhos, robôs que aprendem a andar etc[38].

3.4 Aplicações de *machine learning* na física nuclear

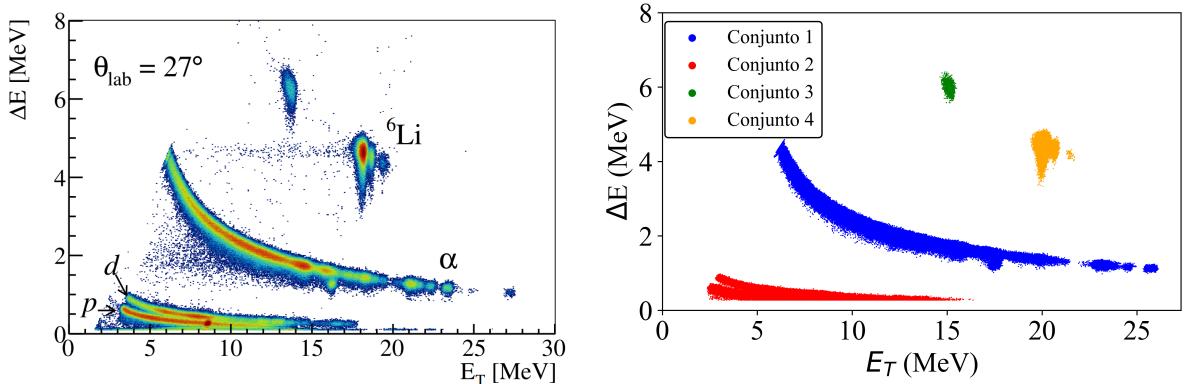
Em física nuclear, o uso de técnicas de *machine learning* tem se mostrado cada vez mais importante. Na continuação são citados alguns exemplos.

3.4.1 Análise de espectros para identificação de partículas (*particle identification, PID*)

O uso de técnicas de aprendizado não supervisionado pode ser usado para a identificar de partículas em espectros $\Delta E - E$. Neste tipo de espectros, as partículas são detectadas e separadas por regiões que facilmente podem ser identificadas visualmente[39]. No entanto, a identificação visual desses espectros com os sistemas de detecção com milhares de canais é inviável de forma manual. Assim, ferramentas de *machine learning* (ou inteligencia artificial) são de grande relevância para esse tipo de analise.

Os resultados apresentados nessa seção foram obtidos na etapa inicial desta dissertação. Para o presente estudo, o espectro biparamétrico ($\Delta E-E$) mostrado na Figura 3.5a foi analisado. Os dados correspondem a reação de espalhamento e transferência no sistema ${}^6\text{Li} + {}^{12}\text{C}$.

O objetivo é identificar conjuntos (ou *clusters*) diferentes a partir do espectro. É claro para o olho humano que existem conjuntos diferentes, e eles podem ser identificados usando algoritmos de aprendizado não supervisionado, como por exemplo o *density-based spatial clustering of applications with noise* (DBSCAN)[40]. O DBSCAN consegue identificar *clusters* apenas com as informações características de densidade existente pelos conjuntos que existem nos dados. O resultado da aplicação do algoritmo está na figura 3.5b, onde é mostrado que o espectro agora está separado por diferentes conjuntos.



(a) Espectro biparamétrico $\Delta E \times E$ (energia no referencial do centro de massa).

(b) Espectro biparamétrico $\Delta E \times E$ (energia no referencial do centro de massa) com os conjuntos identificados com o uso do DBSCAN.

Figura 3.5: Espectros biparamétricos $\Delta E \times E$, em 3.5a o espectro cru com uma pré marcação das partículas, e em 3.5b os conjuntos identificados usando *machine learning*.

Usando o conjunto 1, por exemplo, é possível criar o histograma de identificação de partículas dado pela figura 3.6.

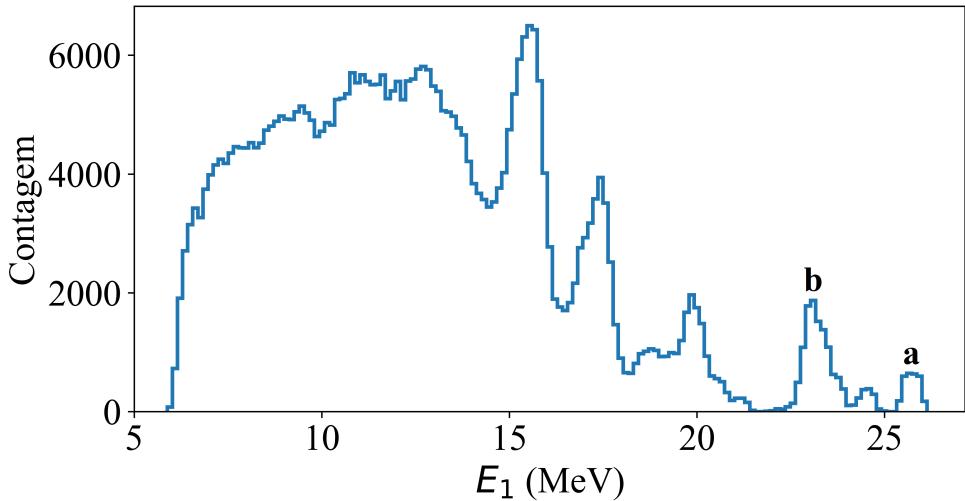


Figura 3.6: Histograma de identificação de partículas, a partir do conjunto 1 identificado na figura 3.5b. Em **a** temos o primeiro pico que corresponde ao estremo fundamental do ^{14}N e em **b** temos o primeiro estado excitado do ^{14}N .

3.4.2 Estimativa de raios e massas nucleares

Frequentemente é necessário calcular com alta precisão observáveis que ainda não foram medidos, para contribuir com dados já existentes. É possível estimar propriedades de núcleos usando modelos e dados experimentais já existentes em conjunto com técnicas de *machine learning*.

É possível usar redes neurais que façam previsão de massa nucleares com informações de massa já disponíveis. Isso foi feito usando o algoritmo *Light Gradient Boosting Machine* (LightGBM)[41]. O algoritmo é uma rede neural cujo aprendizado é supervisionado e minimiza o erro entre a energia de ligação teórica e a energia de ligação experimental dos núcleos (essa diferença é chamada de resíduo), usando 10 quantidades físicas como dados de entrada (*input*)[42]. O desvio quadrático médio para a massa dos núcleos é de 0.234 ± 0.022 MeV, valor melhor que muitos modelos físicos de massa nuclear. A figura 3.7 mostra os resíduos calculados entre a energia de ligação calculada pelo rede neural e a energia de ligação determinada experimentalmente.

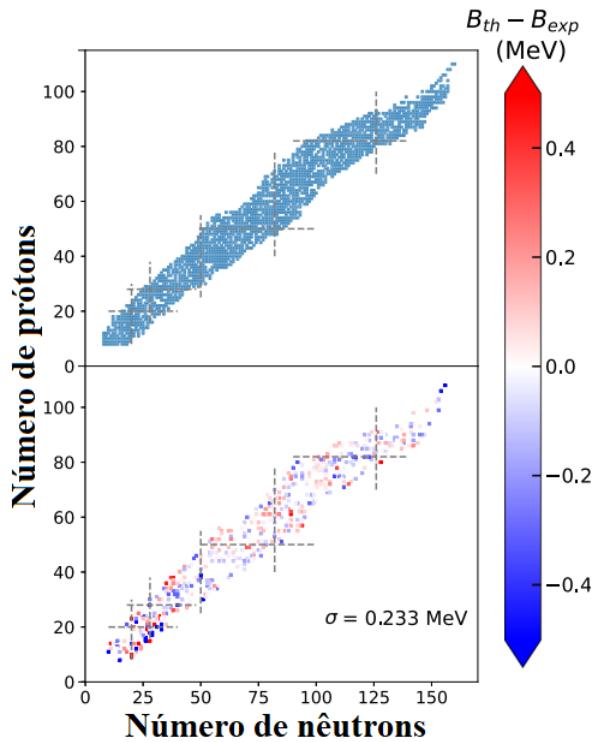


Figura 3.7: Painel acima mostra a localização dos núcleos usados para o treino da rede neural. No painel de baixo temos o erro entre a previsão e o valor experimental da energia de ligação do núcleo para os núcleos usados como dados de validação. σ é o erro quadrático médio da rede neural.

Outro exemplo de uso é para a estimativa de raio de carga nuclear. Usando uma rede neural (chamada de *Bayesian neural network extended liquid drop* - BNN-ELD) FC com duas variáveis de entrada, o número de prótons Z e o número de massa A , com a condição de que $Z \geq 20$ e $A \geq 40$. A rede é supervisionada e usa 722 núcleos para dados de treino e 98 núcleos como validação. Resultados para o desvio quadrático médio do raio de carga são de cerca de 0.02 fm para os 820 núcleos utilizados. A figura 3.8 mostra a diferença entre o raio de carga calculado teoricamente e o determinado experimentalmente para

isótopos do chumbo, usando diferentes neurais e modelos teóricos. Mais detalhes podem ser encontrados na Ref. [43].

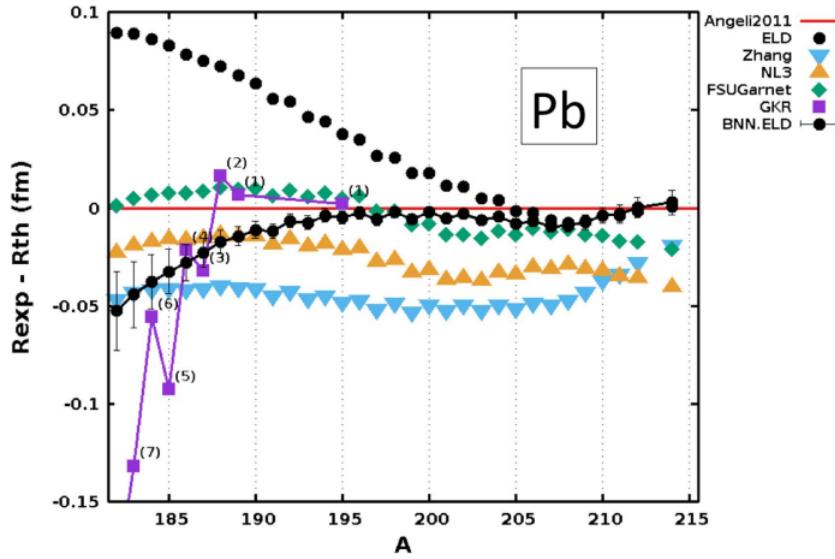


Figura 3.8: Previsões para o raio de carga para isótopos do chumbo ($Z = 82$) para diferentes modelos teóricos ou que usam redes neurais. A previsão que contém barras de erro é a que foi brevemente descrita no texto.

3.4.3 Decaimento β e processo r

O decaimento β é fundamental para entender a origem dos elementos pesados. Prever o tempo de meia vida do decaimento β é de grande importância para simulações do processo r (captura rápida de nêutrons). Com redes neurais é possível fazer previsões que levam em conta a física do problema, como visto na Ref. [44]. O modelo de inteligência artificial leva em conta a teoria de Fermi para o decaimento beta, onde a função f que é minimizada (função custo) é dada por

$$f = \log_{10}(T_{1/2}^a/T_{1/2}^b), \quad (3.7)$$

onde $T_{1/2}^a$ é o tempo de meia vida do decaimento beta medido experimentalmente e $T_{1/2}^b$ é o tempo de meia vida do decaimento beta determinado teoricamente. As variáveis de entrada são o número de prótons Z , o número de nêutrons N , a energia total do decaimento beta e o parâmetro de paridade $\delta = 1, 0, -1$, para núcleos par-par, ímpar-par e ímpar-ímpar, respectivamente.

A figura 3.9 mostra a previsão do tempo de meia vida ($T_{1/2}$) do decaimento β (em segundos) para isótonos com número de nêutrons $N = 126$ usando redes neurais.

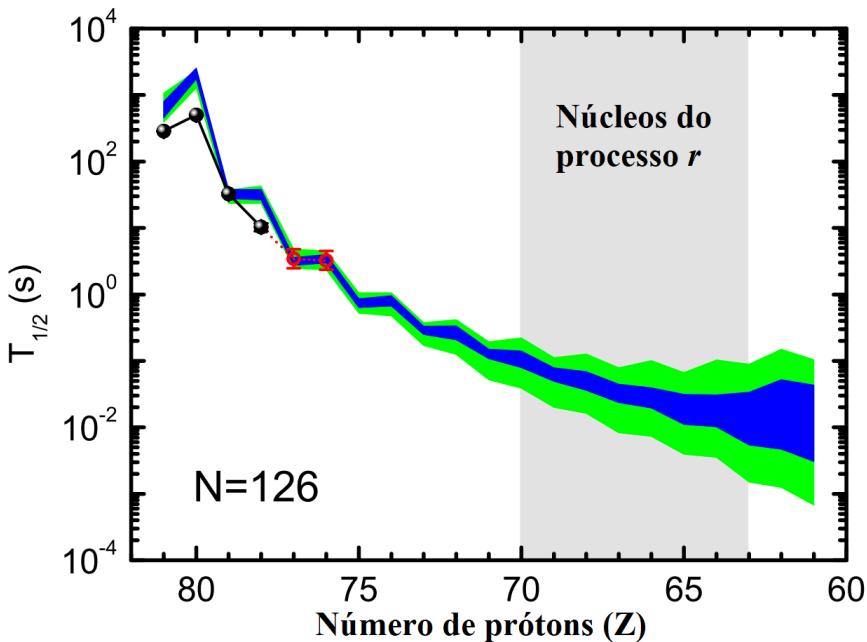


Figura 3.9: Meias-vidas de decaimento β para isótonos com $N = 126$. A região hachurada verde mostra as previsões de uma rede neural. A região hachurada em azul mostra os resultados da mesma rede neural, porém seus dados de aprendizado são estendidos para incluir três meias-vidas extras de decaimento β para cada isótopo (indicado por círculos abertos) em direção à *drip-line* de nêutrons[44].

3.4.4 Alvos ativos

Experimentos com alvos ativos geram enormes quantidades de dados. Uma semana de experimento pode gerar até 10Tb de dados[45], o que gera a necessidade do uso de algoritmos de *machine learning* para diminuir o consumo de tempo da análise desses dados.

A análise dos dados envolve a reconstrução tridimensional dos eventos e posteriormente a reconstrução da cinemática, com a identificação de partículas e de reações nucleares. O uso de técnicas de *machine learning* pode diminuir o consumo de tempo dessas etapas.

O uso de CNNs em imagens feitas a partir das projeções de eventos pode ser útil para a classificação de eventos, sem a necessidade de reconstruir a cinemática em uma etapa anterior. A figura 3.10 mostra exemplos de projeções de trajetórias de partículas dentro do alvo ativo, onde a partir da projeção no plano xy é possível identificar a partícula que a originou.

Primeiro, com o objetivo de classificar as projeções entre próton ou carbono, é possível construir uma rede neural para a classificação binária. Para isso foi usada uma arquitetura

com camadas convolucionais seguidas de *max-pooling* e por fim uma camada FC[45]. A função a ser minimizada é a dada pela equação 3.5 e a métrica para entender o comportamento da rede neural é a acurácia binária.

Caso o objetivo seja classificar entre três ou mais possibilidades (como por exemplo próton, carbono e outros), então a classificação não será binária, passará a ser categórica, pois será necessário classificar a imagem em alguma categoria. A função custo passará a ser a *categorical cross-entropy* e a métrica será a acurácia categórica.

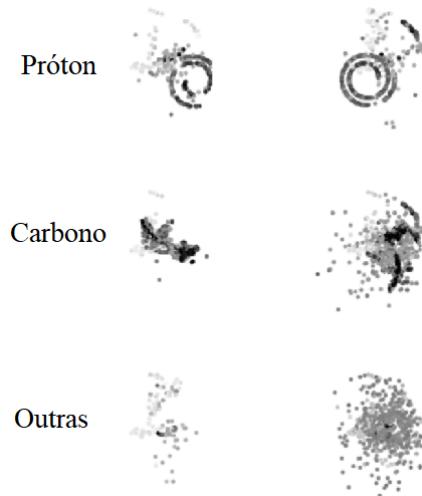


Figura 3.10: Projeções no plano xy de partículas dentro do alvo ativo, onde quanto mais escura for a cor, mais carga tem o ponto, e quanto mais clara a cor, menos carga tem o ponto. O objetivo da rede neural pode ser classificar corretamente se a imagem à direita corresponde à uma trajetória de um próton, carbono ou outra partícula, ou pode ser uma rede neural para classificação binária caso seja necessário classificar apenas entre próton ou carbono[45].

A rede criada para a classificação binária (entre próton e carbono) atingiu cerca de 90% de acurácia. Somado ao fato de que redes neurais são muito eficientes em tempo, isso mostra uma possibilidade de que seja possível utilizar redes neurais como essa para o processamento em tempo real dos dados, com pouca supervisão humana durante o processo[45].

Outro problema que pode ser resolvido com técnicas de *machine learning* é a remoção de ruído de eventos. Para isso, pode ser usado o algoritmo não supervisionado *Hough transform*[46], que é capaz de detectar padrões de interesse nos dados e evitar regiões ou pontos que não seguem o padrão desejado.

Com essa breve descrição sobre *machine learning* e exemplos do seu uso em problemas de física nuclear, pode-se seguir adiante e entender seu uso dentro deste trabalho, que será feito nos próximos capítulos.

Capítulo 4

Reconstrução de nuvens de pontos a partir de algoritmos de *machine learning*

Esse capítulo descreve o procedimento usado para criar as nuvens de pontos (*point-clouds*) a partir dos pulsos gerados por cada pixel do *micromegas*, usando algoritmos de *machine learning* supervisionado.

Algoritmos baseados em CNNs vem sendo usados com relativo sucesso para o processamento e análise de sinais[47], como por exemplo para discriminação de pulsos [48]. CNNs possuem a capacidade de fazer ajustes multidimensionais e aprender padrões extremamente complexos. Além disso, CNNs são mais eficientes em tempo para a análise de grandes quantidades de dados em comparação com algoritmos comuns[47].

A quantidade de pulsos gerados nesse experimento é da ordem de centenas de milhões. Para a análise completa dos pulsos, são necessárias diferentes etapas que envolvem processos que muitas vezes precisam ser refeitos por causa de algum erro no processo de análise. Caso uma etapa dure cerca de 10 horas, notar algum erro no processo, ou ter que mudar algum parâmetro da análise, significaria ter que analisar novamente os dados, sendo um processo muito custoso em tempo.

O uso de algoritmos de CNNs tem o objetivo de diminuir o tempo consumido para a análise dos pulsos. As redes neurais criadas são treinadas uma única vez e podem ser utilizadas de modo separado e/ou acoplado[47], como está mostrado na sequencia do capítulo.

A análise dos dados experimentais foi dividida em três etapas: correção do fundo, deconvolução do sinal e detecção de picos. As redes neurais criadas foram supervisionadas, o que significa que foi preciso determinar a entrada (*input*) e a saída (*output*) das

redes. Para isso foi feito um grande banco de dados, para utilizar no treinamento das redes neurais, que se beneficiam da grande quantidade de dados usada para o treino[16].

Na seção 4.1, está mostrado como o banco de dados para o treino das redes neurais desenvolvidas foi criado, e a construção das redes neurais está na seção 4.2.

4.1 Construção do banco de dados para as redes neurais

Essa seção descreve o processo de construção do banco de dados que foi usado para o treino das redes neurais das subseções 4.2.1, 4.2.2 e 4.2.3. Centenas de sinais foram produzidos pelo detector micromegas em cada evento. Cada um dos pixels do micromegas possui uma eletrônica independente que gera uma resposta como os mostrados na figura 4.1.

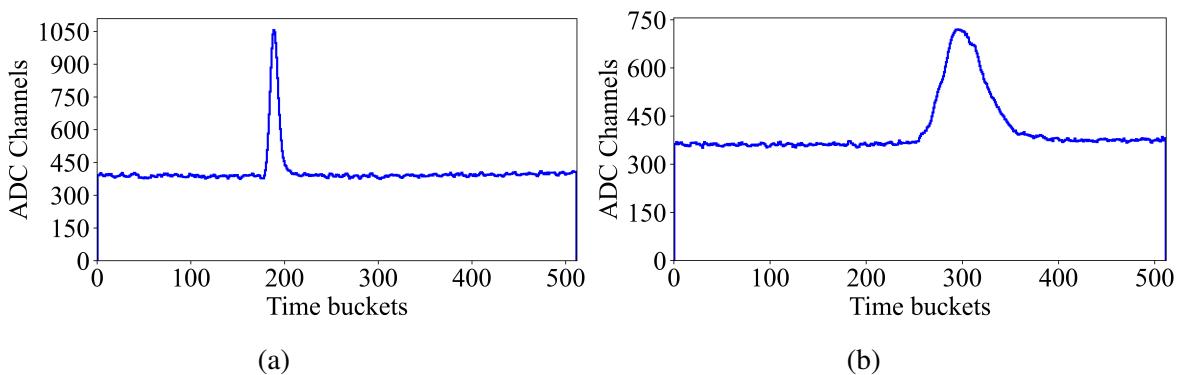


Figura 4.1: Exemplos de sinais produzidos pelos canais do detector. Em 4.1a o sinal possui apenas um pulso, enquanto em 4.1b há vários pulsos em sobreposição, formando um único pulso com largura maior que em 4.1a.

No eixo x , cada um dos 512 *time buckets* possui largura de 195 ns . No eixo y tem-se a carga acumulada no detector para cada *time bucket*. Na figura 4.1a há um sinal com um pedestal (fundo ou *baseline*) com altura entre 300 e 450, e um pulso estreito em cima. Como mostrado na seção 2, os elétrons que surgiram da ionização do gás foram conduzidos perpendicularmente pelo campo elétrico até o detector. A interação da partícula com o gás é evidenciada justamente pelo pulso presente em 4.1a. Cada pixel i do detector está em uma posição (x_i, y_i) , o centroide de cada gaussiana fornece a coordenada em t (*time bucket*) para então ser convertida na posição em coordenada z do ponto de interação da partícula com o gás, e a energia depositada Q é obtida da área do pulso sem fundo (gaussiana com centroide t).

Para a figura 4.1a tem-se apenas um pulso estreito, o que corresponde à um feixe incidindo paralelamente àquele canal do plano detector (perpendicular ao campo elétrico),

pois a projeção da interação da partícula com gás no tempo é uma distribuição estreita. No caso da figura 4.1b, há uma distribuição ampla do sinal do tempo, o que corresponde ao feixe incidindo perpendicularmente ao canal do plano detector. A ilustração desse processo está na figura 4.2, que mostra o processo da passagem de uma partícula carregada e como o sinal é gerado a partir disso.

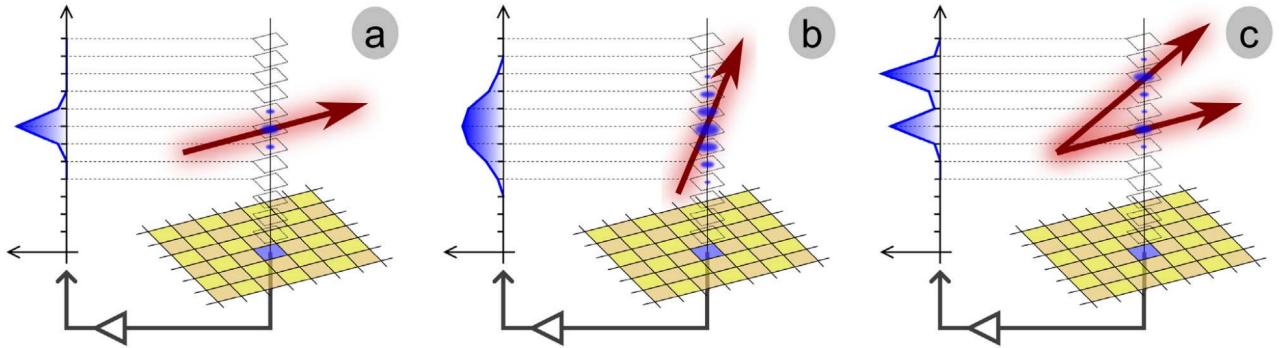


Figura 4.2: Ilustração que mostra a variação no formato da carga coletada a partir da passagem de uma partícula carregada dentro do TPC, onde o plano do detector está embaixo. No lado esquerdo de cada imagem, a distribuição do sinal coletado por um único pad (escuro) do plano de coleta é mostrado (o canal eletrônico de leitura é representado pela seta cinza em negrito). No caso de uma trajetória quase horizontal em relação ao plano do detector (a), o sinal é uma distribuição estreita, enquanto para uma trajetória próxima a uma direção vertical (ou perpendicular) em relação ao detector (b), a distribuição deve ser muito mais ampla (vários pontos de interação da partícula com o gás devem ser extraídos desse sinal). A última imagem ilustra o caso em mais de uma trajetória de partículas contribui para o sinal[14].

Para analisar os pulsos deve-se primeiro remover o fundo (pedestal ou *baseline*) dos sinal. O sinal de fundo é complexo e pode variar por canal e também por evento. Desde flutuações causadas pelo circuito eletrônico até efeitos sistemáticos gerados pela memória de *buffer* circular alteram o sinal, podendo o tornar não analítico [47, 14]. Com o sinal sem o fundo, se faz a deconvolução para determinar todos os centroides e cargas acumuladas dos pontos, para a reconstrução da nuvem de pontos. Todo esse processo foi realizado nesse trabalho com algoritmos de *machine learning* supervisionado[47]. Para isso, foi criado um banco de dados que serviu de *output* e/ou *input* para o treino das redes neurais.

A criação dos dados para o treino das redes neural está mostrado na seção 4.1.1. A criação das redes neurais está mostrada na seção 4.2.

4.1.1 Estimativa do fundo

Essa subseção descreve a estimativa do fundo de cada, para formar o banco de dados da rede neural que estima o fundo descrita na subseção 4.2.1.

A primeira tentativa de estimar o sinal sem o fundo é usando transformada de Fourier e um filtro passa-baixa, que é a função resposta do detector fornecida na Ref.[14]. Seja $f(t)$ uma função qualquer, sua transformada de Fourier é dada por

$$\hat{f}(\nu) = \mathcal{F}[f(t)] = \int_{-\infty}^{\infty} f(t) e^{-2\pi i \nu t} dt. \quad (4.1)$$

Primeiro calculamos a transformada de Fourier $\hat{f}(\nu)$ do sinal, em seguida multiplicamos pela função resposta do detector $h(\nu)$ dada por[14]

$$h(\nu) = A * \exp(\nu\tau) (\nu\tau)^3 \sin(\nu\tau), \quad (4.2)$$

onde A está relacionado com o ganho de amplificação e τ é o tempo de pico (*peaking time*), que é o tempo de modelagem da cadeia de amplificação [14]. Do teorema da convolução, temos que [49]

$$\mathcal{F}^{-1}[\hat{f}(\nu)\hat{g}(\nu)] = (f * g)(t) = \int_{-\infty}^{\infty} f(\tau)g(t - \tau)d\tau, \quad (4.3)$$

onde $(f * g)(t)$ é a convolução entre $f(t)$ e $g(t)$. Multiplicar o sinal transformado por $h(\nu)$ e depois inverter a transformação é o mesmo que convoluir o sinal original com a transformação inversa de $h(\nu)$, o que resulta no sinal sem o fundo[13, 14]. Resultados desse procedimento estão na figura 4.3.

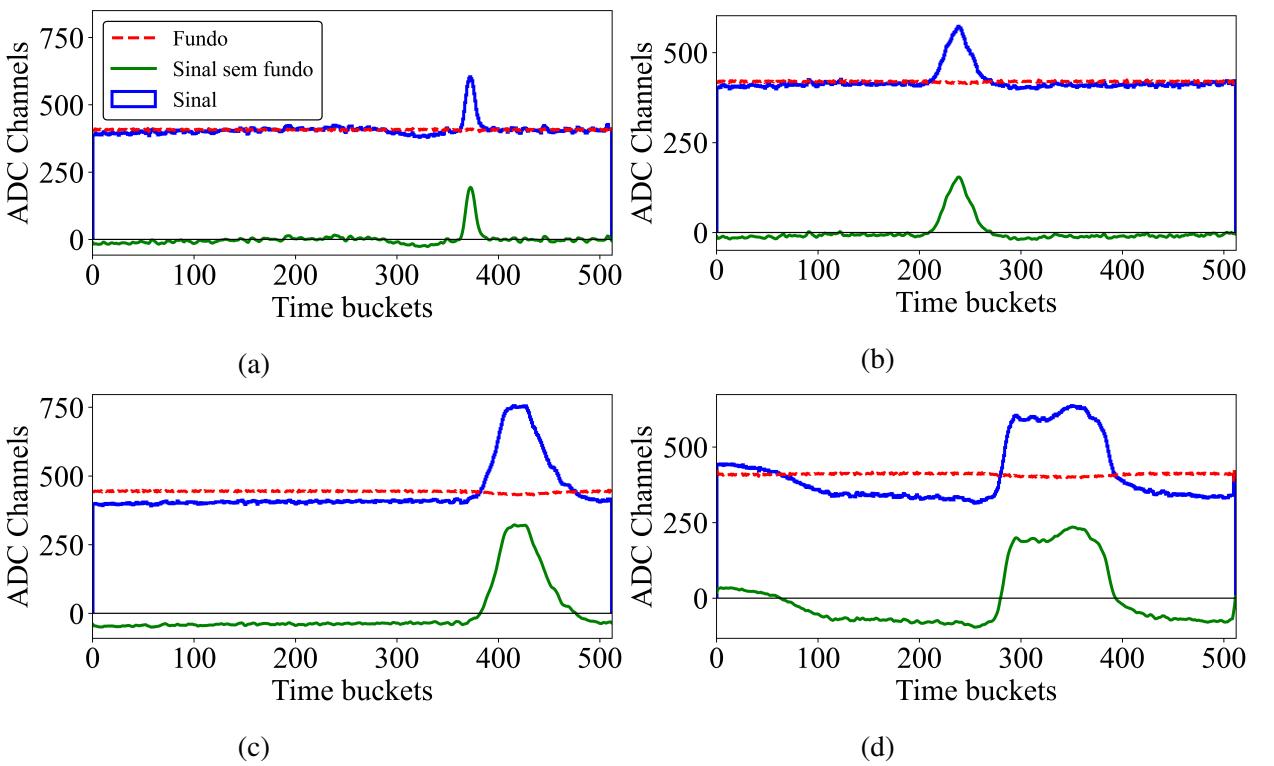


Figura 4.3: Histogramas com as respectivas *baselines* (linhas tracejadas) estimadas pelo método da convolução. O espectro resultante (sem o fundo) está em verde.

Fica claro que visualmente, por exemplo na figura 4.3d, que o filtro utilizado não é a melhor função resposta do detector. Poderia-se estimar essa função resposta empiricamente, porém os canais auxiliares chamados de *Fixed Pattern Noise* (FPN)[14] usados para esta estimativa não foram armazenados. Portanto, a estimativa do fundo foi feita sinal por sinal[47, 14]. Para isso, o fundo foi determinado usando o algoritmo *background removal* da biblioteca *TSpectrum* do *ROOT* [50]. A função tem a capacidade de separar o fundo dos picos presentes no espectro[51, 52, 53]. Exemplos de estimativa do fundo estão na figura 4.4.

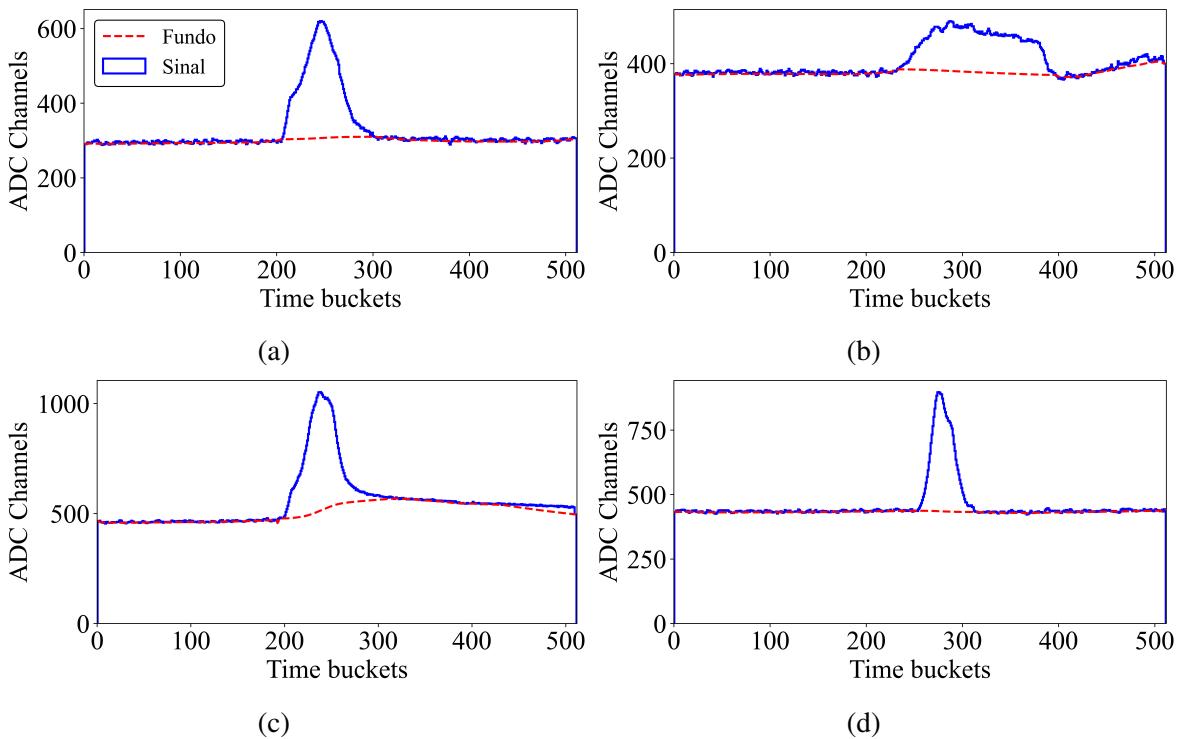


Figura 4.4: Histogramas com as respectivas *baselines* (linhas tracejadas) calculadas pelo *TSpectrum*.

Os resultados do cálculo do fundo de cada sinal foram armazenados e usados para a etapa de deconvolução do sinal, que deve ser feita com o sinal sem o fundo. Para diminuir as flutuações numéricas após a remoção do fundo, o valor mínimo do sinal sem o fundo é zero. Sem o fundo podemos buscar por todos os picos e suas cargas correspondentes no sinal. Não podemos detectar diretamente todos os picos pois muitos deles estão em sobreposição. Para isso foi feita a deconvolução do sinal, descrita na subseção 4.1.2.

4.1.2 Deconvolução do sinal

Para aumentar a resolução dos picos foi usado o algoritmo *gold deconvolution* presente na biblioteca *TSpectrum* do *ROOT*[54]. O algoritmo tem como objetivo fazer a deconvolução do espectro, gerando uma função (nesse caso um sinal) resposta de acordo com o sigma esperado para os pulsos. O sinal resposta corresponde ao espectro com as gaussianas não sobrepostas. Isso significa que foi necessário determinar qual o valor de sigma dos pulsos para buscar a função resposta.

A largura dos pulsos é o mesmo de um sinal que possui apenas um pico. Ou seja, a largura dos pulsos foi determinada fazendo a análise de sinais que possuem apenas 1 pico, fazendo um ajuste pelo método dos mínimos quadrados (MMQ) de uma gaussiana.

Para buscar espectros com apenas um pico, foi usado o algoritmo de detecção de picos *peak_finder*, presente na biblioteca do *scipy*[55], e para o ajuste da gaussiana foi usada o pacote *lmfit* [56]. O valor de sigma encontrado foi de 4.09 (17) *time buckets*.

A largura escolhida foi ligeiramente maior pois, verificando empiricamente, em alguns casos o algoritmo separava o que deveria ser uma única gaussiana em duas. O valor de sigma usado na deconvolução foi de 4.30 *time buckets*. Foi determinado também o número de iterações do algoritmo de deconvolução. O número de iterações escolhido foi de 700, menos que isso o algoritmo não estava separando totalmente picos sobrepostos. O limiar para a escolha de um ponto como um pico foi definido como ter altura maior que 20% do valor máximo do sinal. Resultados da deconvolução estão na figura 4.5.

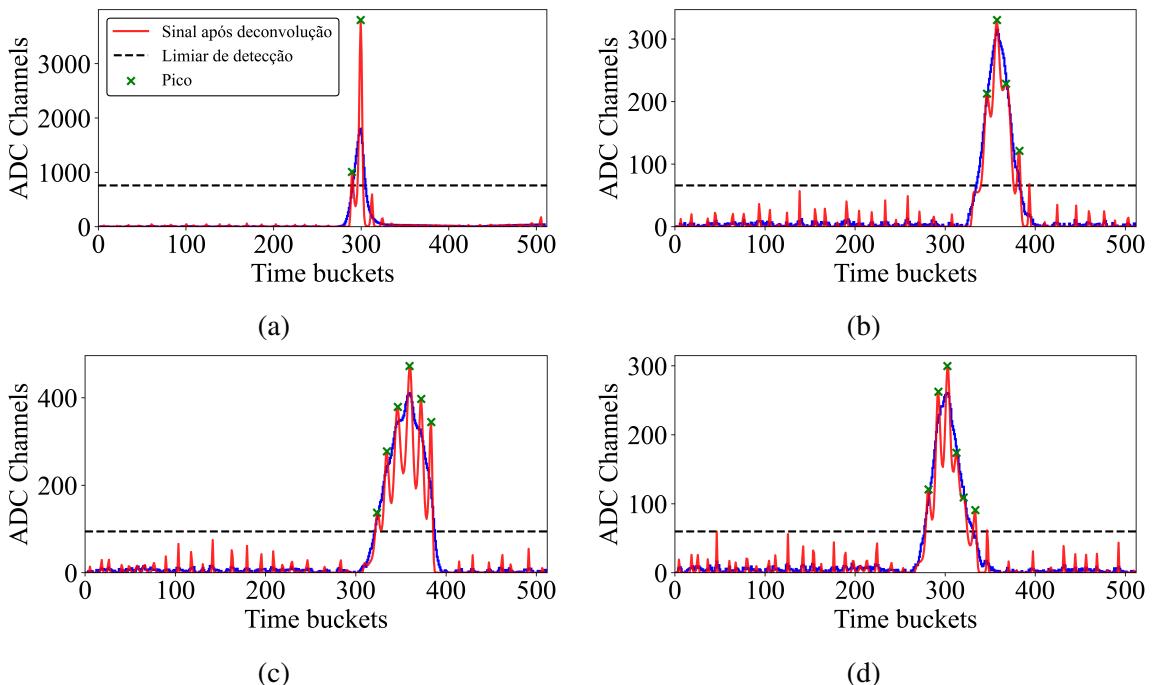


Figura 4.5: Histogramas sem as *baselines* antes (em azul) e depois da deconvolução (em vermelho). Os picos (em verde) e o limiar (linha tracejada preta) de detecção também estão indicados.

O algoritmo de deconvolução também retorna a posição dos centroides encontrados, que indica a localização de um pico. Os mesmos picos podem ser obtidos com o *peak_finder*, com a vantagem de que o algoritmo possui muitos parâmetros diferentes para calibração, melhorando a detecção em comparação com os picos detectados pelo algoritmo do TSpectrum. A execução de 200.000 sinais, desde a estimativa e remoção do fundo, até a detecção dos centroides, demora cerca de 23.25 minutos, usando o processador Ryzen 5 3600X.

Para determinar a carga acumulada Q de cada ponto, é necessário calcular a área do

centroide do pico detectado. A área do sinal antes e depois da deconvolução é a mesma, mesmo para a região dos pulsos, portanto pode-se analisar diretamente o sinal após a deconvolução. Para achar a área do pulso, foi calculada a largura dos pulsos após a deconvolução, para determinar a área como uma simples integral gaussiana. O sigma dos pulsos após a deconvolução é $\sigma_{dd} = 1.1543$ (44) *time buckets*. Com isso foi calculada a carga acumulada para cada ponto descoberto do evento. A carga acumulada Q para cada ponto i é dada por:

$$Q = \int_{-\infty}^{\infty} Ae^{-(t'-t_i)^2/2\sigma_{dd}^2} dt' = A |\sigma_{dd}| \sqrt{2\pi}, \quad (4.4)$$

onde A é a amplitude do ponto com centroide t_i e desvio padrão após a deconvolução σ_{dd} .

Com o banco de dados para a deconvolução e também para a detecção de picos, basta criar as redes neurais, que serão descritas na seção 4.2.

4.2 Análise dos pulsos com *machine learning*

Com *machine learning* temos a possibilidade de criar algoritmos de alta complexidade sem definir operações explícitas. Usando os resultados das seções anteriores foram desenvolvidas três redes neurais, com o objetivo de: estimar o fundo (subseção 4.2.1), fazer a deconvolução (4.2.2) e por fim detectar os picos (subseção 4.2.3).

4.2.1 Rede neural para o fundo

O objetivo foi criar uma rede neural que reproduza o comportamento do algoritmo *background removal* que estima o fundo do sinal, que foi discutido na seção 4.1.1, tentando reproduzir resultados muito similares. A rede neural é supervisionada, onde os dados de entrada são os sinais brutos e as saídas devem ser os fundos de cada sinal. A arquitetura está na figura 4.6.

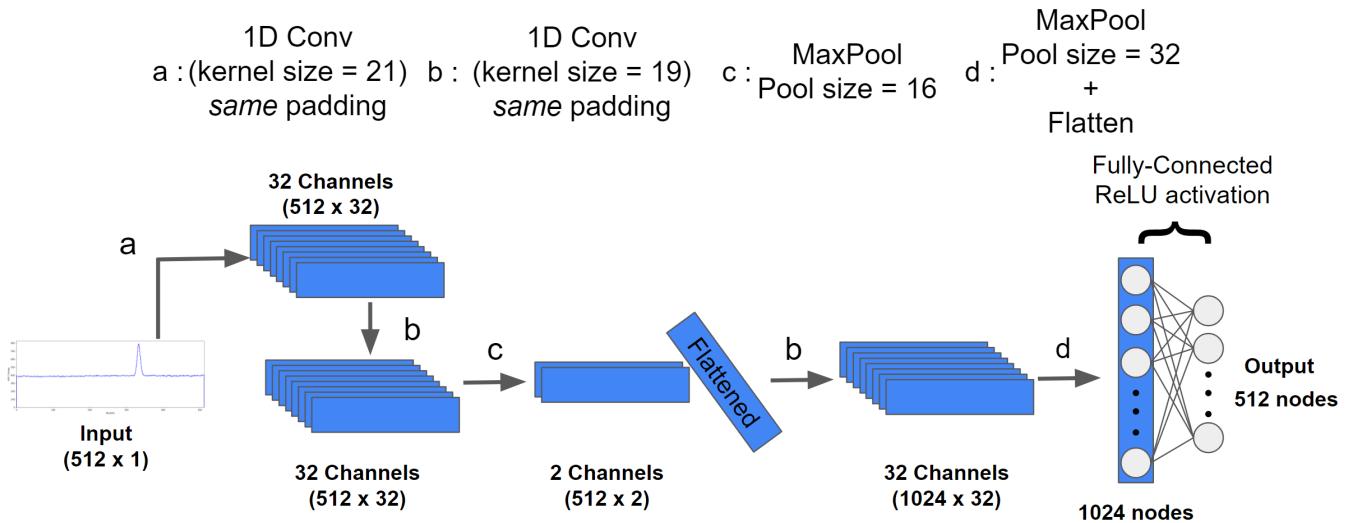


Figura 4.6: Arquitetura da rede neural que faz a inferência do fundo. O vetor de entrada deve ter dimensionalidade 512×1 . Todas as partes com convolução não possuem o parâmetro *bias*.

A entrada da rede é o sinal cru com dimensionalidade 512×1 . Há duas convoluções seguidas (passagens *a* e *b*) com *padding same*, seguida de uma camada com *Max pooling*. Os dois canais restantes sofrem uma planificação (ou *flat*), diminuindo sua dimensionalidade, para então passar por mais uma convolução com *padding same* e filtros de tamanho 19 seguido de uma camada *Max pooling* e uma camada *Fully Connected* com função de ativação *ReLU*. Toda a rede neural foi construída usando o TensorFlow 2 e possui um total de 545.536 parâmetros, todos treináveis. O tamanho dos filtros das convoluções levam em conta a largura do pulso, sendo no mínimo maior que a largura, a fim de que cada *kernel* atue em um pulso completo na convolução[47].

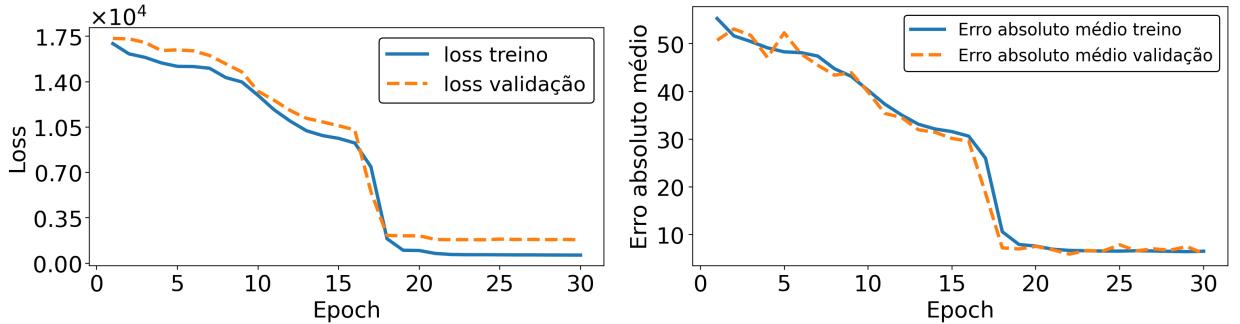
A camada final com função de ativação *ReLU* garante o valor mínimo de saída em 0 e, principalmente, pelo fato de não causar problemas à minimização do gradiente [57]. Foram testadas diversas combinações e a mostrada na figura 4.6 é a que obteve os melhores resultados[47].

Para o treino foram usados 160.000 sinais para treino e 40.000 para validação. O *loss* foi escolhido como sendo o erro quadrático médio (equação 4.5, o otimizador foi o ADAMAX [30], com *learning rate* de 0.0005, e métrica para avaliação foi o erro médio absoluto, dado por

$$E = \frac{1}{N} \sum_{i=1}^N |x_i - \hat{x}_i|, \quad (4.5)$$

onde E é o erro absoluto médio, N é o número de pontos e x_i o ponto da saída da rede para ser comparado com o ponto original \hat{x}_i . Foram 30 *epochs* e o *batch-size* foi 8.

O treino foi realizado no Google Colaboratory [58] usando a GPU (*graphics processing unit*) NVIDIA Tesla P100 e durou cerca de 34 minutos. Os resultados do treino estão na figura 4.7.



(a) *Loss* dos dados de treino (linha contínua) e dos dados de validação (linha tracejada) em função da *epoch* no treino da rede dada pela figura 4.6.

(b) Erro absoluto médio dos dados de treino (linha contínua) e dos dados de validação (linha tracejada) em função da *epoch* no treino da rede dada pela figura 4.6.

Figura 4.7: Resultados do treino da rede neural dada pela figura 4.6. A rede atingiu seu melhor resultado a partir da *epoch* 20 aproximadamente, quando começa um platô no *loss*.

A arquitetura da figura 4.6 (assim como as próximas desse capítulo) foi determinada de forma empírica. Uma arquitetura com menos passagens e /ou menos parâmetros fornece resultados menos adequados em comparação com a arquitetura apresentada. No caso de mais parâmetros e /ou passagens (consequentemente com aumento no tempo de execução), a rede neural não demonstrou melhora substancial.

Exemplos de resultados de previsões da rede neural estão na figura 4.8. A previsão do fundo possui um erro absoluto nos dados de treino de 6.5315 ADC Channels e nos dados de validação de 6.0783 ADC Channels. O sinal cru é subtraído do fundo, colocando o valor mínimo da subtração em 0. Comparando o erro médio absoluto de 200.000 sinais sem o respectivo fundo (resultante do algoritmo do TSpectrum) com o resultado da rede neural obtemos 4.5 ADC Channels.

Rede neurais convolucionais têm a vantagem de usarem poucas variáveis e serem facilmente paralelizadas em sua execução[16]. Uma vantagem de redes neurais é o seu tempo de execução. Empiricamente a rede neural pode processar 200.000 sinais em apenas 8s (ou 25.000 sinais por segundo), sendo muito eficiente em tempo.

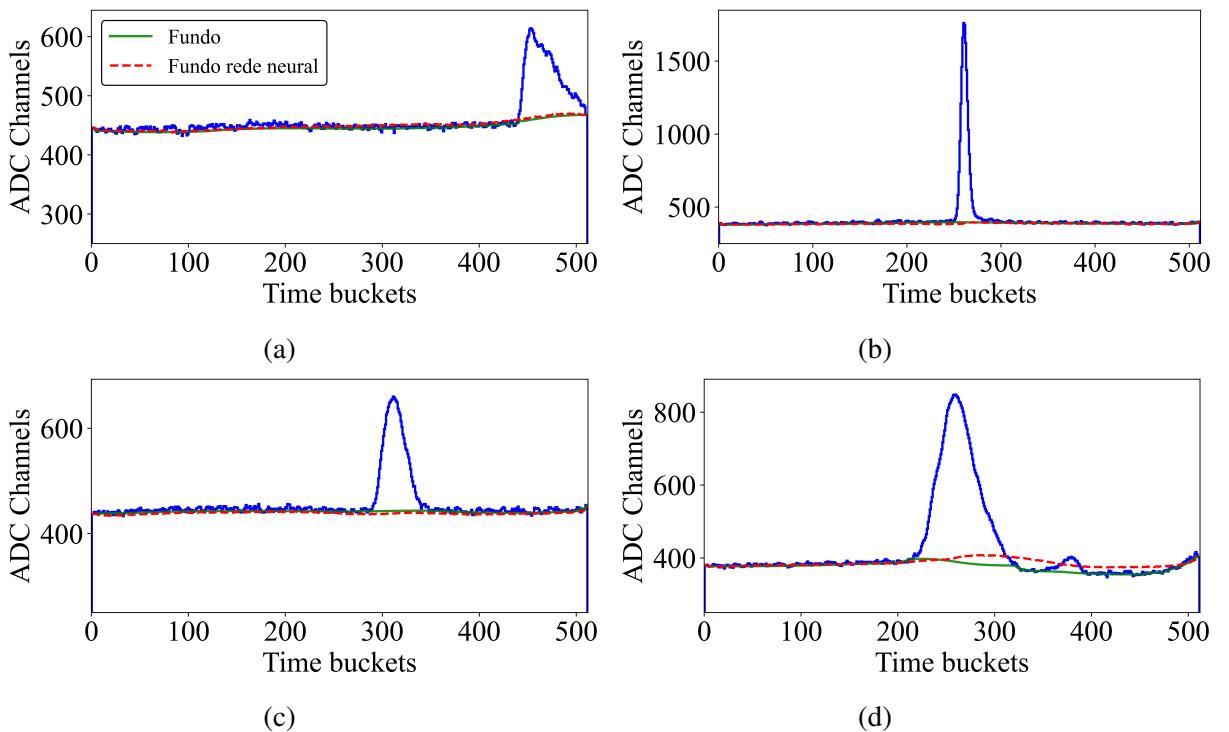


Figura 4.8: Exemplos da rede neural dada pela figura 4.6 em comparação com a saída do *TSpectrum*.

Nos exemplos mostrados nas figuras 4.8a, 4.8b e 4.8a os fundos dos sinais possuem grande flutuação e a rede neural se mostrou eficaz na previsão. No exemplo 4.8d a *baseline* do sinal é muito complexa de se determinar, pois o sinal, aproximadamente do canal 300 a 500, varia em cerca de 50 unidades em y . Apesar da rede neural determinar o fundo acima do fundo original, ao subtrair o espectro do fundo e colocar o valor mínimo em 0, o pulso presente entre os canais 200 e 300 é praticamente inalterado.

Com os resultados obtidos pela rede neural que calcula o sinal de fundo, o próximo passo foi criar a rede neural que faz a deconvolução do espectro sem a *baseline*.

4.2.2 Rede neural para a deconvolução

A mesma abordagem da rede neural anterior foi usada, que é fazer uma sequencia de convoluções e por fim uma camada *fully connected* com função de ativação *ReLU*, pois precisamos ter o valor mínimo do espectro em 0. Os filtros das convoluções precisam ter tamanho mínimo de duas vezes o sigma das gaussianas para atuarem sobre cada pulso do espectro. A entrada da rede é o sinal com o fundo subtraído e com mínimo em 0. A saída é o sinal após deconvolução dada pelo algoritmo *gold deconvolution* na biblioteca *TSpectrum* do ROOT, já mostrado na subseção 4.1.2. A figura 4.9 mostra a arquitetura da

rede de deconvolução.

A rede é a sequência de duas convoluções com 32 filtros, *valid padding* e *kernels* de tamanho 19 e 17, respectivamente, seguida de uma camada *Max pooling* com *pool size* igual à 16. No final há o *flat* na camada para seguir com uma camada *fully connected* com função de ativação *ReLU*. O *valid padding* se mostrou mais eficiente para a convergência da rede. Toda a rede foi construída usando o TensorFlow 2, possuindo 508.000 parâmetros treináveis[47].

Assim como na rede anterior, foram usados 160.000 sinais para treino e 40.000 para validação. O *loss* foi escolhido como sendo o erro quadrático médio, o otimizador foi o *ADAM*, com *learning rate* de 0.0005 porém com o parâmetro *clipnorm* igual a 0.45. A métrica para avaliação foi o erro médio absoluto. Foram 75 *epochs* e o *batch-size* foi 8. Os resultados do treino estão na figura 4.10.

Alterar a norma do gradiente (usar o parâmetro *clipnorm* = 0.45) significa que, caso a norma do vetor do gradiente exceda 0.45, então o valor da norma é reajustado para o limiar (*threshold*) escolhido (0.45)[47]. Isso faz com que não ocorra problemas comuns como o gradiente sumir[57, 30], o que estava acontecendo especificamente nesse caso.

O treino foi realizado no Google Colaboratory [58] usando a GPU NVIDIA Tesla P100 e demorou cerca de 54 minutos. Os resultados do treino estão na figura 4.7, onde eles indicam que, visualmente, a rede neural consegue distinguir muito bem diferentes centroides presentes no pulso. Empiricamente, a rede é capaz de executar 200.000 sinais em 5.4 segundos (ou 37.000 sinais por segundo).

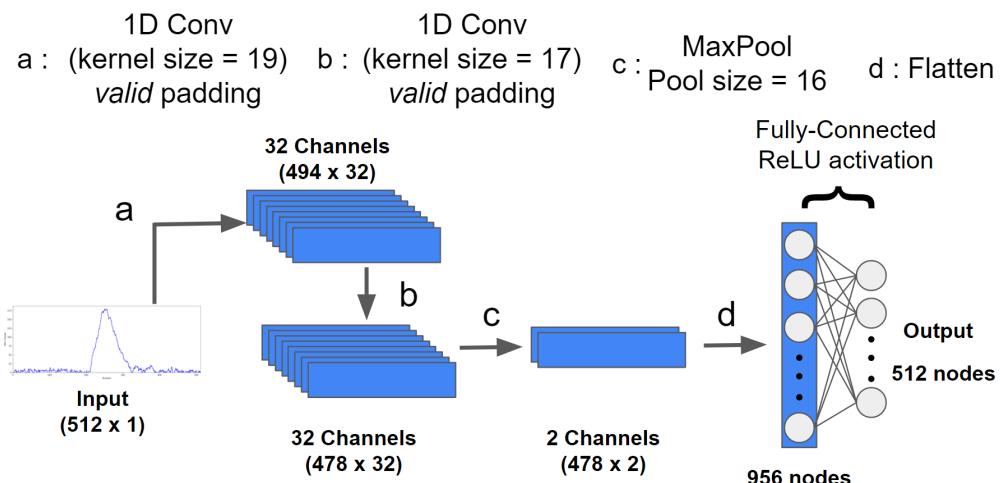
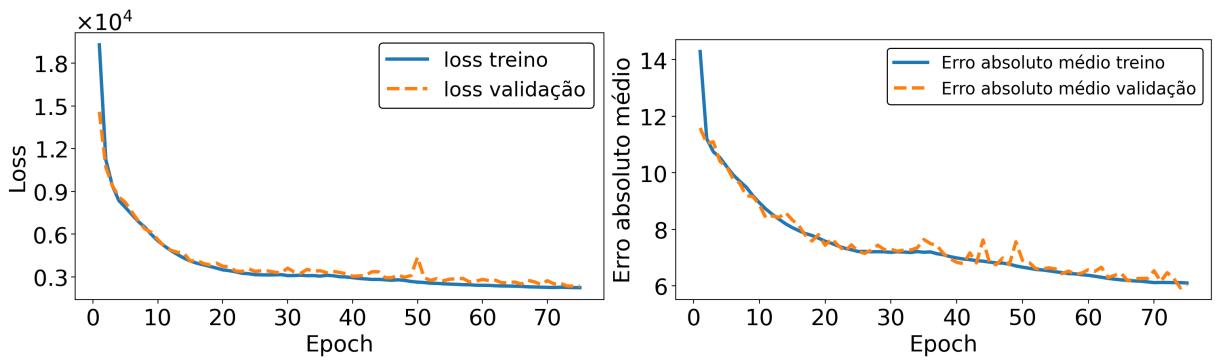


Figura 4.9: Arquitetura da rede neural que faz a inferência da deconvolução do espectro. O vetor de entrada deve ter dimensionalidade 512 x 1. Todas as partes com convolução não possuem o parâmetro *bias*.



(a) *Loss* dos dados de treino (linha contínua) e dos dados de validação (linha tracejada) em função da *epoch* no treino da rede dada pela figura 4.9.

(b) Erro absoluto médio dos dados de treino (linha contínua) e dos dados de validação (linha tracejada) em função da *epoch* no treino da rede dada pela figura 4.9.

Figura 4.10: Resultados do treino da rede neural dada pela figura 4.6.

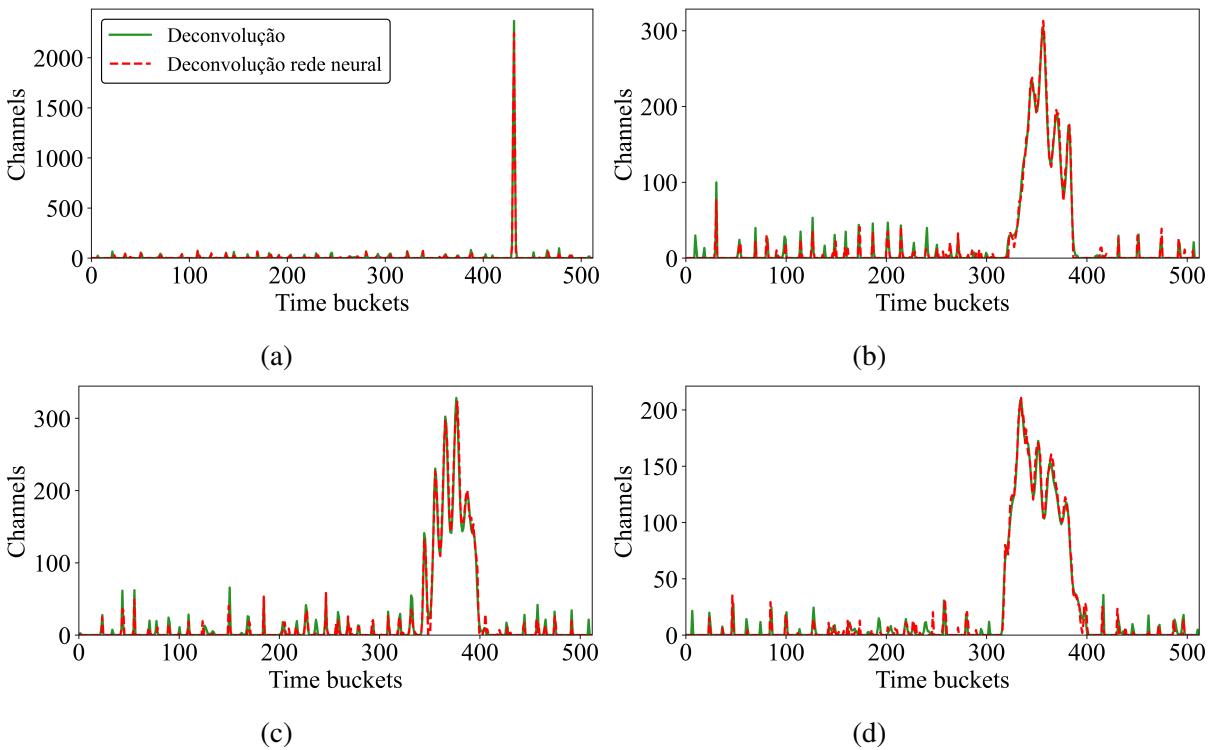


Figura 4.11: Exemplos de deconvolução da rede neural dada pela figura 4.6.

Com a rede neural para a deconvolução feita, a última etapa é a construção de uma rede neural para a detecção de picos, mostrada na 4.2.3.

4.2.3 Detecção de picos

A última etapa dessa análise com *machine learning* foi analisar possíveis soluções para a detecção de picos. Esse é um problema muito complexo, pois há uma variação muito grande na quantidade de picos por sinal e também há um desbalanço muito grande na quantidade de pontos comuns (aqueles que não são picos) e pontos que são picos. Por exemplo, caso haja um sinal que possui apenas um pico, devemos detectar uma posição, dar o valor de saída como 1, por exemplo, dentre 512 pontos, onde 511 terão o valor de saída como 0. Caso a rede determine que todos os pontos são não picos, ainda assim a acurácia binária seria maior que 99%. Isso é conhecido como desbalanço de classe[59].

Para corrigir esse desbalanço, foram acrescentados pontos simetricamente em torno do pulso, de forma que a somatória das amplitudes das regiões é equivalente a carga Q acumulada no ponto. Isso faz com que não seja preciso detectar um único ponto, mas sim uma região em torno do pico. Com isso podemos nos basear na ideia de segmentar o sinal para destacar regiões de interesse[60]. Segmentar significa ter uma rede neural com a saída com o mesmo tamanho do vetor de entrada (512) e saída com valores entre 0 e 1, onde 1 indica uma região com um pico e 0 não. A figura 4.12 mostra um exemplo de um sinal após a deconvolução onde há os picos detectados com o algoritmo *peak_finder* e os pontos acrescentados simetricamente em torno dos picos para representar as regiões dos pulsos.

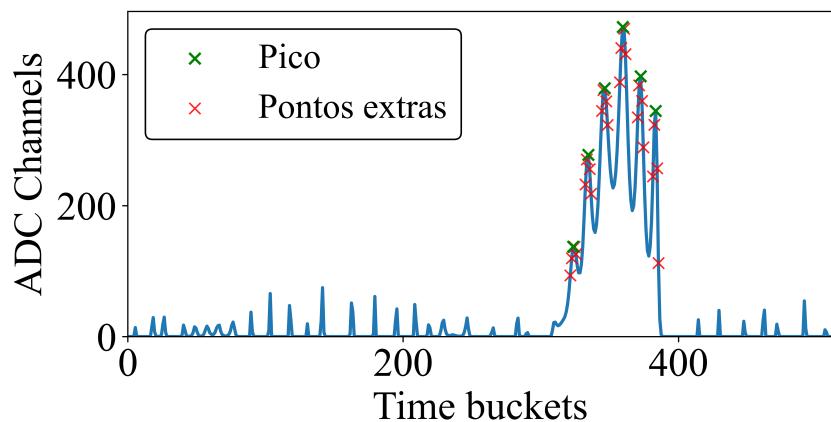


Figura 4.12: Sinal após a deconvolução que mostra o pico detectado mais os pontos adicionais que irão facilitar o trabalho da rede neural (evitar o desbalanço de classe). Foram acrescentados 2 pontos à esquerda e à direita.

A rede construída é a sequência de uma convolução com *kernel* de tamanho 13 e *same padding* seguida de *Max-Pooling* e uma *fully-connected* com função de ativação sigmoide, possuindo um total de 263.104 parâmetros treináveis. Para o treino foram usados sinais

que possuíam entre 1 e 6 picos, resultantes da saída do algoritmo *peak_finder* do *SciPy*, o que resultou em 120.024 de dados para o treino e 30.006 para validação. A escolha pelos picos detectados pelo *peak_finder* ao invés do algoritmo de detecção do TSpectrum é pela maior flexibilidade de ajuste fino do algoritmo, tornando a detecção de picos muito melhor[47]. A função custo escolhida foi a *binary cross-entropy* (dada pela equação 3.5), o otimizador o *ADAM* com *learning rate* de 0.001. A métrica utilizada foi a acurácia binária. O treino também foi realizado por uma GPU NVIDIA Tesla P100 e durou cerca de 8 minutos com 12 *epochs*. A arquitetura da rede está na figura 4.13 e os resultados do treino estão na figura 4.14.

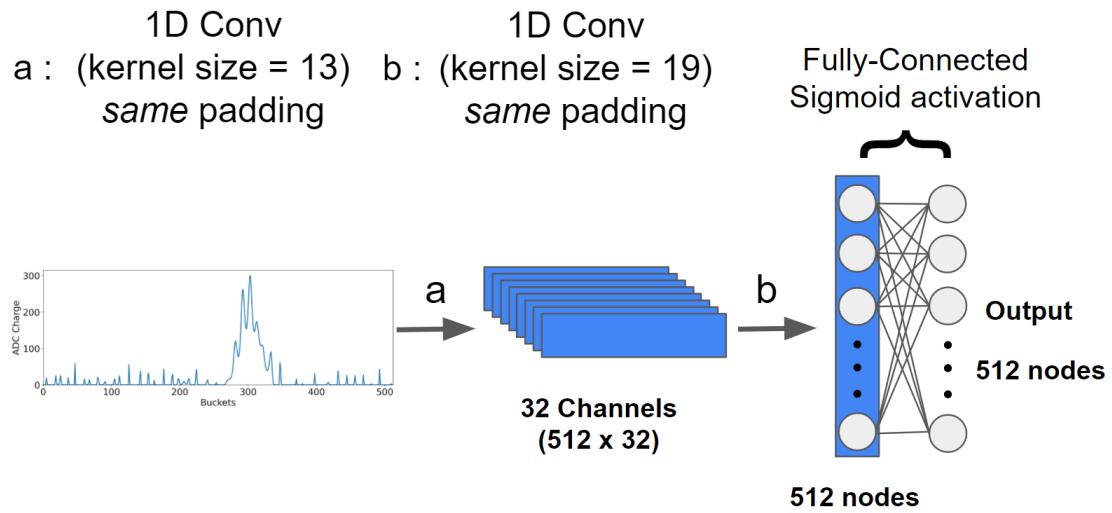


Figura 4.13: Arquitetura da rede neural que faz o recorte das regiões com picos. O vetor de entrada deve ter dimensionalidade 512 x 1.

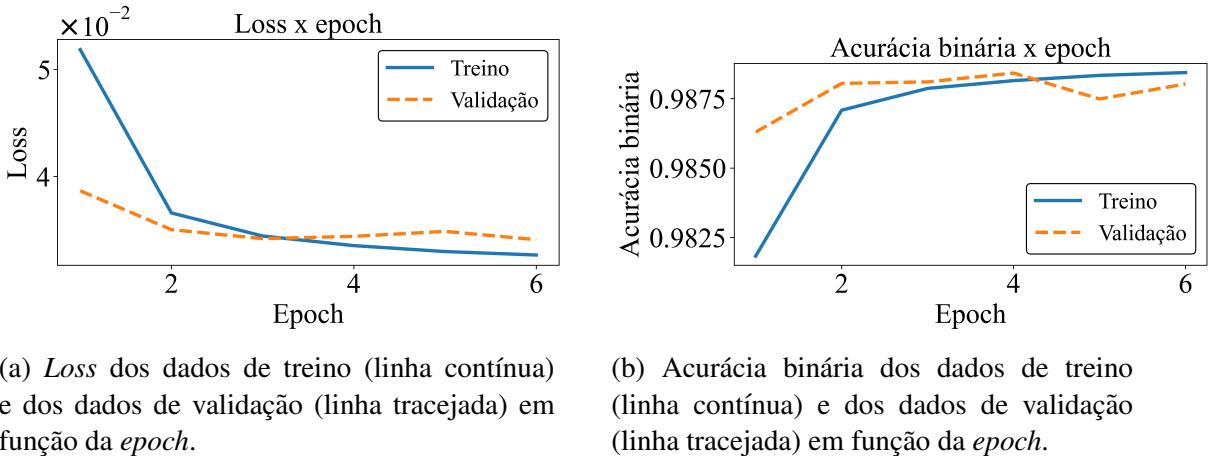


Figura 4.14: Resultados do treino da rede neural dada pela figura 4.13.

A saída da rede neural é um vetor de tamanho 512 com valores entre 0 e 1, onde valores maiores que 0.5 são considerados como pertencentes à um pulso. Com as regiões

identificadas podemos fazer uma média ponderada com o espectro de entrada para achar o centroide. O tempo de processamento da rede neural é de 150.030 sinais em cerca de 4.11 segundos (aproximadamente 36.500 sinais por segundo). Para determinar os picos a partir da saída da rede neural, o tempo é de aproximadamente 4.3 segundos, onde o algoritmo pode ser ainda mais rápido se for paralelizado. Resultados para picos detectados pela rede neural em comparação com o algoritmo *peak_finder* estão na figura 4.15.

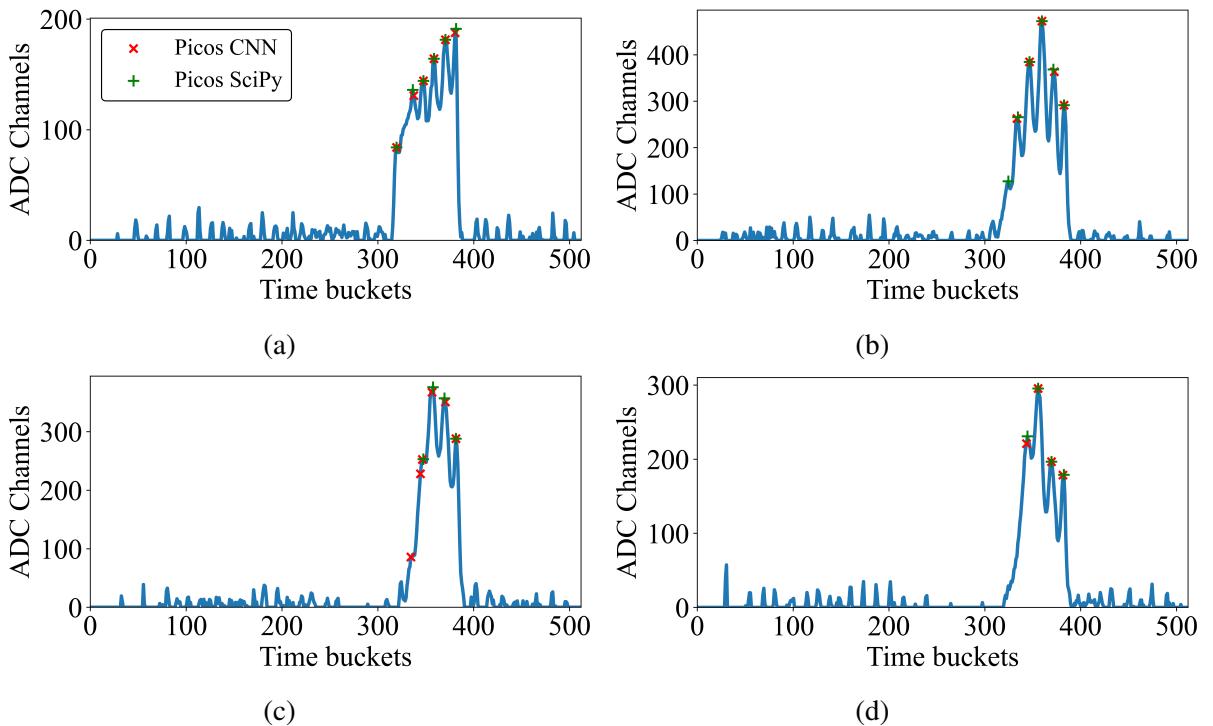


Figura 4.15: Exemplos de detecção de picos usando a rede neural, em comparação com a detecção feita pelo algoritmo presente no SciPy, mostrada na figura 4.13. Os centroides detectados pela rede neural estão em vermelho, e os centroides detectados pelo SciPy estão em verde. Em azul está o espectro sem fundo após a deconvolução, resultantes

Para determinar carga acumulada Q_i de cada ponto associado à cada centroide i , foi usada a relação

$$Q_i = \sum_{i-2}^{i+2} f(t_i) \alpha_\sigma, \quad (4.6)$$

onde $f(t_i)$ é a amplitude do espectro no *time bucket* t da posição i e α_σ é uma constante real para calibrar o valor da área em função do sigma dos pulsos, que foi determinado empiricamente como $\alpha_\sigma = 1.2$.

Com as três redes neurais criadas, é necessário verificar o acoplamento das redes, a fim de analisar a qualidade dos algoritmos e verificar os tempos de execução, para comparar

com os algoritmos discutidos nas subseções 4.1.1 e 4.1.2.

4.2.4 Acoplando as redes neurais

Usando novamente o TensorFlow 2 pode-se carregar as redes neurais discutidas nas subseções 4.2.1, 4.2.2 e 4.2.3, já treinadas e usar como se fosse uma única rede neural. Acoplando as três arquiteturas, temos a nova arquitetura mostrada na figura 4.16. O *input* da rede é o sinal cru o *output* da rede neural é o um vetor de tamanho 1024, que possui a segmentação e o sinal após a deconvolução, pois as duas informações são necessárias para determinar os centroides.

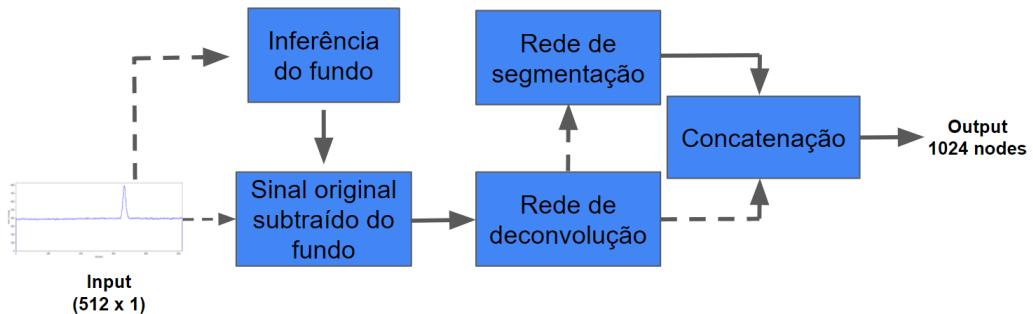


Figura 4.16: Arquitetura da rede neural que faz a inferência da *baseline*, em seguida faz a deconvolução do espectro sem o fundo e por fim faz a segmentação do sinal. O resultado da segmentação e da deconvolução são concatenados na parte final da rede neural. O vetor de entrada deve ter dimensionalidade 512 x 1.

A rede é apenas a sequencia das redes anteriores, ou seja, o espectro passa pelo cálculo da *baseline*, então o espectro original é subtraído dessa *baseline* (colocando o valor mínimo em 0) para passar pela etapa da deconvolução e em seguida o sinal é segmentado. Na etapa final, o resultado da segmentação e da deconvolução são concatenados em um único vetor, para assim determinar os centroides. Pelo fato de cada parte ser treinada de modo separado não há necessidade de treinar a rede unificada, apenas carregar as variáveis das redes neurais treinadas de cada parte. A rede unificada possui 1.316.640 de parâmetros.

Foi comparada a saída da rede acoplada para o espectro após a deconvolução, com a saída de referência, que é o espectro após a deconvolução mostrado na subseção 4.1.2. Podemos usar novamente o erro médio absoluto para fins de comparação. O erro médio absoluto, para os 200.000 sinais, é de 7.45 ADC Channels. Com relação à incerteza, ela foi estimada como o desvio padrão da diferença de cada *time bucket* do sinal tido como referência e o sinal após a rede neural acoplada. A incerteza é da ordem de 6% da amplitude do sinal no ponto.

Com relação a eficiência em tempo, a rede neural da figura 4.16 processa 200.000 sinais em 12 segundos, usando a GPU NVIDIA Tesla P100. Somando esse tempo com a determinação dos centroides, que é de aproximadamente 4.3 segundos, o tempo total para processar 200 mil sinais é de cerca de 16.3 segundos, cerca de 90 vezes mais rápido que os métodos mostrados na seção 4.1. Exemplos da reconstrução das nuvens de pontos usando a rede neural da figura 4.16 são apresentados na figura 4.17d.

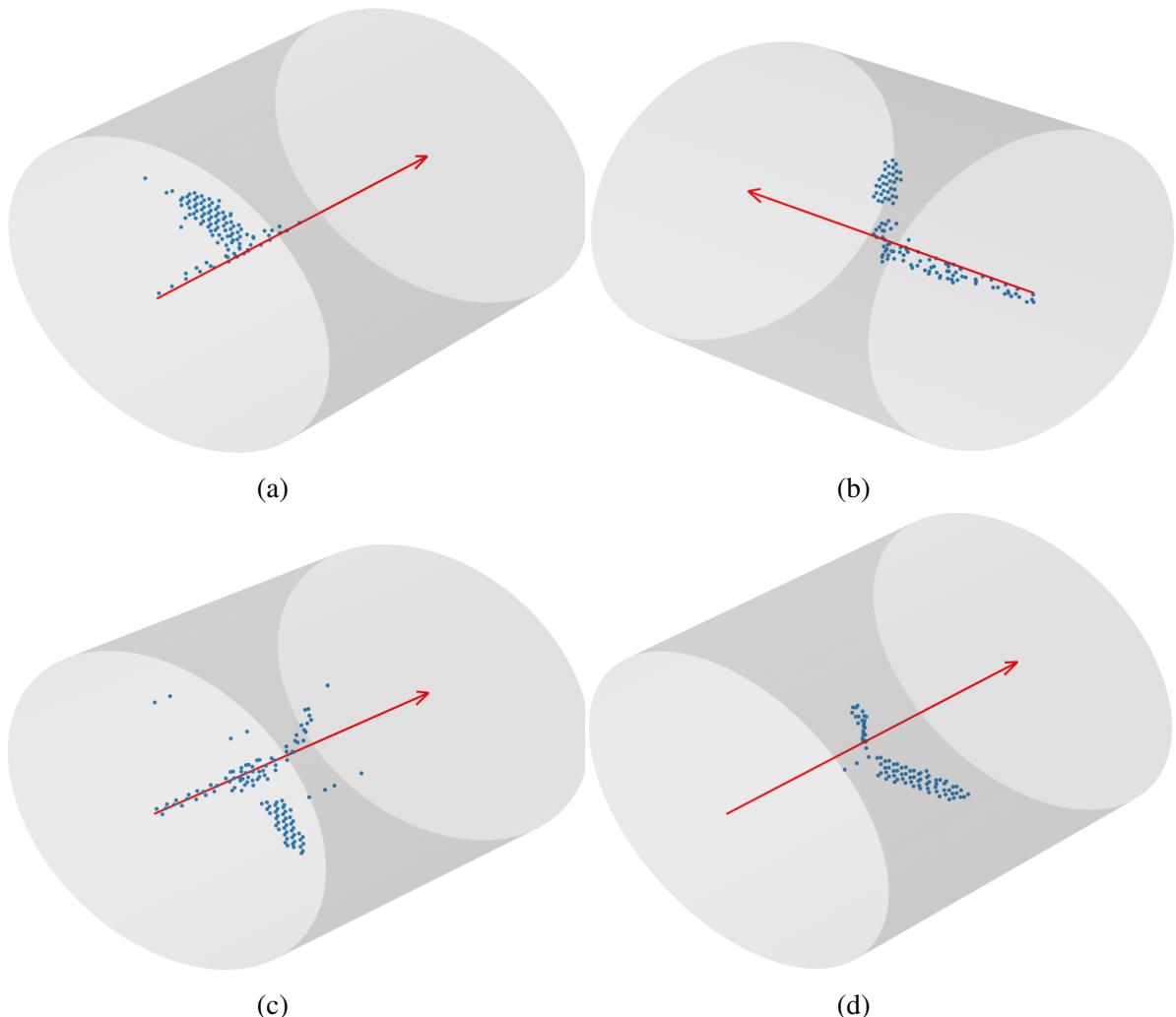


Figura 4.17: Exemplos de eventos reconstruídos através da análise dos sinais com *machine learning*. A seta vermelha indica o sentido do feixe.

Com as redes neurais desenvolvidas, o consumo de tempo para o processamento dos pulsos diminuiu em cerca de 90 vezes em relação à algoritmos mais tradicionais mostrados na seção 4.1, abrindo possibilidades para a análise em tempo real de um experimento. Com as nuvens de pontos reconstruídas, pode-se extrair propriedades físicas dos eventos, processo que será descrito no capítulo 5.

Capítulo 5

Análise das nuvens de pontos

Esse capítulo irá mostrar como foram analisadas as nuvens de pontos com algoritmos de *machine learning*. O objetivo desta etapa, do ponto de vista computacional, foi de detectar *clusters*, que nesse caso são retas tridimensionais. Já do ponto de vista físico, o objetivo é, com as trajetórias identificadas, extrair as informações físicas (energia E , momento \vec{p} e comprimento L). Com isso, foi possível distinguir cada reta detectada, e determinar o vértice de reação (entre uma partícula originada da reação nuclear entre o feixe e o gás). Após a detecção, foram selecionados os eventos que possuíam reações nucleares.

Assim como no capítulo 5, esse capítulo descreve o uso de algoritmos de *machine learning* para a análise, desta vez para as nuvens de pontos. Os algoritmos usados foram tanto não supervisionados (algoritmos de *clustering*) e supervisionados (redes neurais), com diferentes objetivos.

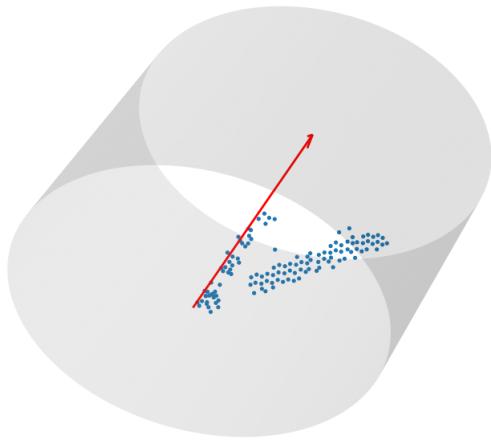
Na seção 5.1 está mostrado o processo de detecção de trajetórias (*tracking*) com algoritmo de *machine learning* não supervisionados. Na seção 5.2.1 estão mostradas alternativas com outros algoritmos de *machine learning*, porém que não foram utilizadas para a análise dos dados, apenas estudadas suas respectivas viabilidades.

5.1 Detecção de trajetórias

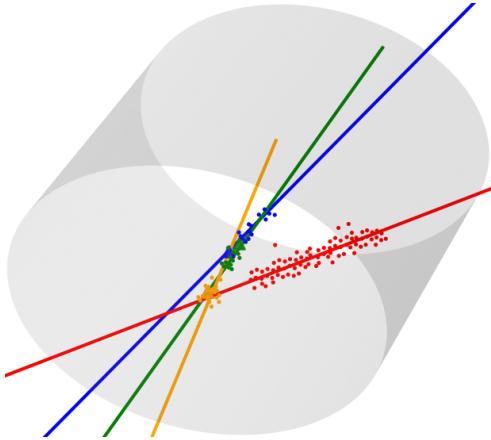
Essa seção irá descrever o processo de *tracking* usando algoritmos de *machine learning* não supervisionados. As nuvens de pontos (eventos) que foram analisadas são como o da figura 5.1a. Nessa figura é notável ao olho humano que os pontos formam estruturas de retas. Essa retas foram detectadas com algoritmos de *clustering*, pelo fato da estrutura aparente dos dados e, também, por não existir um tamanho fixo das nuvens de pontos. O número de pontos pode variar de 30 até cerca de 200, não considerando pontos ruidosos.

Para diminuir a quantidade de pontos ruidosos, foram usados dois filtros. O primeiro filtro exclui pontos baseado na sua carga, colocando um limiar onde pontos com carga $Q < 110$ são descartados. O segundo filtro, chamado de *outlier removal*, elimina pontos considerados *outliers globais*, de modo que, caso um ponto não possua um número mínimo de vizinhos $n_{or} = 4$ em um raio de distância $d_{or} = 12$ mm, então ele é descartado. O *outlier removal* está presente na biblioteca Open3D[61] no Python e funciona excluindo pontos muito isolados uns dos outros.

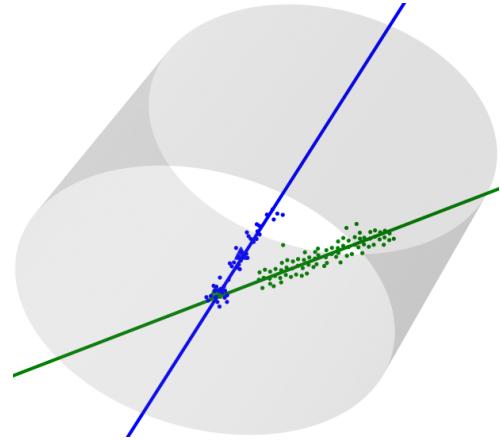
Existem diversas opções de algoritmos de clustering, como o *Density-based spatial clustering of applications with noise* (DBSCAN)[40], porém a escolha deve levar em conta a performance em tempo, por isso o algoritmo usado foi o *Hierarchical DBSCAN* (HDBSCAN) [62, 63]. O algoritmo tem como entrada a nuvem de pontos e parâmetros esperados para a densidade dos *clusters* presentes no conjunto de pontos. O algoritmo retorna os diferentes *clusters* identificados. Os *clusters* então passam por um ajuste por mínimos quadrados, para determinar o versor e um ponto arbitrário que determinam a reta tridimensional. A figura 5.1b mostra o resultado da aplicação do HDBSCAN na nuvem de pontos mostrada na figura 5.1a. O algoritmo não é perfeito e possui falhas em alguns resultados, como por exemplo na figura [xx]. As nuvens de pontos reconstruídas não possuem, em certos casos, densidade de pontos o suficiente para resultar em 100% de acurácia do evento.



(a) Exemplo de evento analisado. Os pontos em azul são das partículas detectadas pelo TPC, a seta vermelha indicando o sentido do feixe e o TPC está representado pelo cilindro cinza.



(b) Evento com as detecções sem a correção. As três retas de cores amarela, verde e azul são de um único *cluster*.



(c) Evento corrigido. Agora o evento possui as duas retas corretas, a azul e a verde.

Figura 5.1: Sequência de análise de um evento. Em 5.1a temos o evento que é recebido para ser analisado, em 5.1b temos o mesmo evento após o HDBSCAN (antes da correção) e 5.1c mostra depois da correção. As cores das retas são arbitrárias e servem apenas para a diferenciação.

O HDBSCAN apresenta falhas no seu resultado, como visto na figura 5.1b, em que, por exemplo, um único *cluster* acaba sendo dividido em três *clusters* muito próximos. A próxima etapa foi de correção da saída do *clustering*. A ideia é comparar, dois a dois, todos os *clusters* resultantes do algoritmo e, se satisfazer um determinado critério, unificar os dois clusters[64]. Do ponto de vista computacional, esse problema é abordado avaliando a semelhança entre dois *clusters* usando métricas, como a distância de Jaccard[65] e o coeficiente de silhueta[66]. A correção feita se dá em duas etapas: primeiro comparando os versores entre duas retas e depois verificando se a condição da equação 5.1 é satisfeita.

Caso a diferença absoluta entre os ângulos com relação ao versor $(0, 0, 1)$ seja menor

que 9° (determinado empiricamente), então as duas retas serão combinadas se obedecerem a condição dada por

$$\sum_{i=0}^{N_1} \frac{d_{i2}}{N_1} < \alpha d_{min}, \quad (5.1)$$

onde N_1 é o número de pontos da reta 1, d_{i2} a distância do ponto i da reta 1 em relação à reta 2, α é um parâmetro com valor a ser escolhido e d_{min} é a distância mínima do ponto a reta. Os valores foram determinados empiricamente, tais que $\alpha = 1.75$ e $d_{min} = 15$ mm. A figura 5.1c mostra o resultado da correção baseada nesses critérios no resultado anterior mostrado na figura 5.1b.

Após a correção, é necessário classificar cada reta como sendo ou o feixe, ou uma partícula originada de uma reação nuclear. O feixe incide na câmara com um ângulo muito pequeno com relação ao versor $(0, 0, 1)$. Além disso, mesmo se o ângulo for muito pequeno, a reta do feixe cruza o plano da janela do TPC muito próximo do ponto mais provável da entrada o feixe. O ponto mais provável foi calculado usando a posição média da projeção dos pontos de um conjunto de eventos no plano x - y . Disso obtemos que a posição inicial mais provável do feixe é tal que $x_f = -3.4$ (6.7) mm e $y_f = -0.9$ (6.3) mm. A incerteza é alta pois o feixe incide de modo bem distribuído da abertura da janela.

Portanto se o ângulo entre o versor \hat{v}_i de uma reta r_i for menor que 5° (determinado de modo empírico novamente) e a distância d entre o ponto P_i que intercepta o plano e o ponto $(x_f, y_f, 0)$, for menor que 15 mm (pouco mais que duas vezes a incerteza de cada ponto), então a reta foi considerada como o feixe do evento. No caso de não satisfazer essas condições, então ela foi classificada como uma possível partícula originada da reação do feixe com o gás.

Importante notar que há eventos que não possuem o feixe detectado, como mostrado na figura 5.2. Neste caso, foi necessário assumir as propriedades da reta mais provável para o feixe, ou seja, precisa passar pelo ponto $(x_f, y_f, 0)$ e ter versor $(0, 0, 1)$.

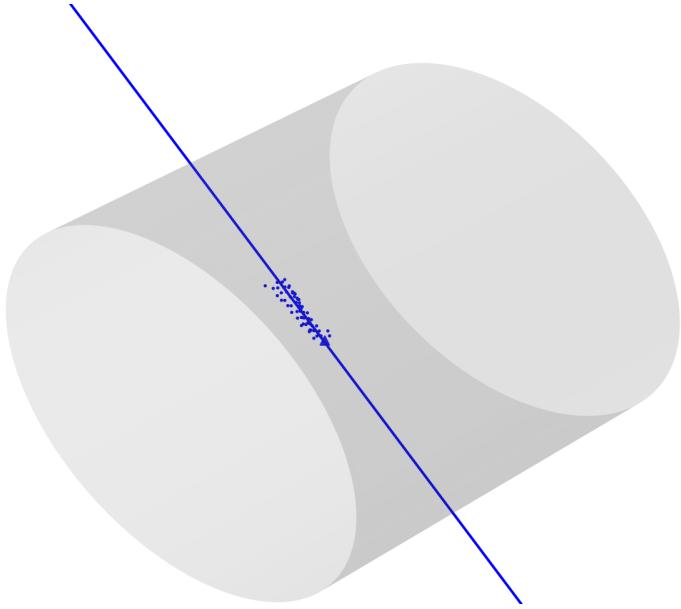


Figura 5.2: Evento em que não foi detectado o feixe, apenas a partícula espalhada. O triângulo azul é o local calculo do vértice de reação dado pela equação 5.6.

Para completar a cinemática do evento, foi necessário calcular o vértice de reação, para cada evento, entre a partícula e o feixe. O vértice de reação é o ponto médio do segmento de reta que conecta a reta a ser analisada (partícula) e o feixe, no ponto de menor distância entre as retas. Ele permitiu fazer correções em etapas futuras e diferenciar o começo e o fim de uma trajetória.

Para deduzir a equação para o vértice de reação, primeiro temos as seguintes equações das retas \vec{P}_1 e \vec{P}_2 como vetores:

$$\begin{aligned}\vec{P}_1 &= \vec{A}_1 + \vec{V}_1 * t_1 \\ \vec{P}_2 &= \vec{A}_2 + \vec{V}_2 * t_2,\end{aligned}\tag{5.2}$$

onde \vec{A}_1 e \vec{A}_2 são pontos arbitrários que pertencem as retas 1 e 2, respectivamente, \vec{V}_1 e \vec{V}_2 são os versores, t_1 e t_2 são os hiperparâmetros das retas.

A reta que conecta a menor distância possui versor

$$\vec{V}_c = \frac{\vec{V}_1 \times \vec{V}_2}{|\vec{V}_1 \times \vec{V}_2|}.\tag{5.3}$$

Podemos então construir uma reta \vec{P}_3 que conecta \vec{P}_1 e \vec{P}_2 . Essa reta deve começar no ponto de menor distância da reta 1 e terminar no ponto de menor distância da reta 2. Ou seja, tem-se o seguinte sistema linear:

$$\vec{A}_2 + \vec{V}_2 * \tilde{t}_2 = \vec{A}_1 + \vec{V}_1 * \tilde{t}_1 + \vec{V}_c * \tilde{t}_3.$$

Rearranjando temos que

$$\vec{V}_1 * \tilde{t}_1 - \vec{V}_2 * \tilde{t}_2 + \vec{V}_c * \tilde{t}_3 = \vec{A}_2 - \vec{A}_1, \quad (5.4)$$

onde \tilde{t}_1 , \tilde{t}_2 e \tilde{t}_3 são os hiperparâmetros a serem determinados. Caso \vec{V}_1 seja paralelo à \vec{V}_2 , então não há solução (não há vértice de reação, ou seja, não há uma reação nuclear em comum entre a partícula e o feixe analisado). Com a solução do sistema pode-se obter os pontos de menor distância nas duas retas:

$$\begin{aligned} \vec{P}_1 &= \vec{A}_1 + \vec{V}_1 * \tilde{t}_1 \\ \vec{P}_2 &= \vec{A}_2 + \vec{V}_2 * \tilde{t}_2. \end{aligned} \quad (5.5)$$

Com isso, é possível determinar que o vértice de reação \vec{V}_r é dado por

$$\vec{V}_r = \frac{1}{2}(\vec{P}_1 + \vec{P}_2). \quad (5.6)$$

Com isso pode-se definir a distância de máxima aproximação d_{max} das retas, dada pela equação 5.7.

$$d_{max} = |\vec{P}_1 - \vec{P}_2|. \quad (5.7)$$

$$d_{max} = \left| (\vec{P}_1 - \vec{P}_2) \right|. \quad (5.8)$$

Da equação 5.7 foi estabelecido um limite superior para d_{max} tal que valores maiores que esse limite indicam uma reação nuclear em comum entre a partícula e o feixe. O valor foi determinado empiricamente e foi definido como $d_{max}^{sup} = 25$ mm. Trajetórias cuja distância máxima de aproximação excedia d_{max}^{sup} , então a trajetória é descartada. A última condição para garantir que houve uma reação nuclear, é garantir que o vértice de reação está dentro da câmara, cujos limites são $|x| < 140$ mm, $|y| < 140$ mm e $|t| < 512$.

Para completar essa etapa, foi necessário determinar a energia E da trajetória e o comprimento L da trajetória. A energia E é definida como

$$E = \sum_{i=1}^N Q_i, \quad (5.9)$$

onde Q_i é a carga acumulada do i-ésimo ponto de um conjunto com N pontos. O comprimento é definido como

$$L = \max d_{i,\vec{V}_r}, \quad (5.10)$$

onde d_{i,\vec{V}_r} é distância entre o ponto i pertencente a trajetória até o vértice de reação \vec{V}_r .

Essa não foi a única abordagem utilizada no trabalho. Na seção 5.2 foram mostradas alternativas para se buscar, com as nuvens de pontos crua, eventos específicos que ocorreram no experimento, sem ter que previamente buscar trajetórias e também outros usos de *machine learning* para a análise.

5.2 Abordagens alternativas

O objetivo da análise das nuvens de pontos foi buscar eventos que possuíam, de modo geral, três trajetórias (uma sendo o feixe, e as outras duas de partículas que surgiram da reação nuclear do feixe com o gás). Nem todos os eventos puderam ser analisados, devido à falta de trajetórias, o que significa que foi utilizado tempo para a busca desses eventos que poderiam ser descartados.

Para facilitar a análise, pode ser necessário investigar eventos atípicos, ou mesmo

5.2.1 Detecção de eventos com *machine learning*

Para evitar o processo de *clustering* de nuvens de pontos que não possuem eventos de interesse (como o *breakup*), é possível usar redes neurais supervisionadas capazes de processar nuvens de pontos tridimensionais para selecionar eventos de interesse. A rede neural usada para esse processo chama-se PointNet [67] e a arquitetura está na figura 5.3.

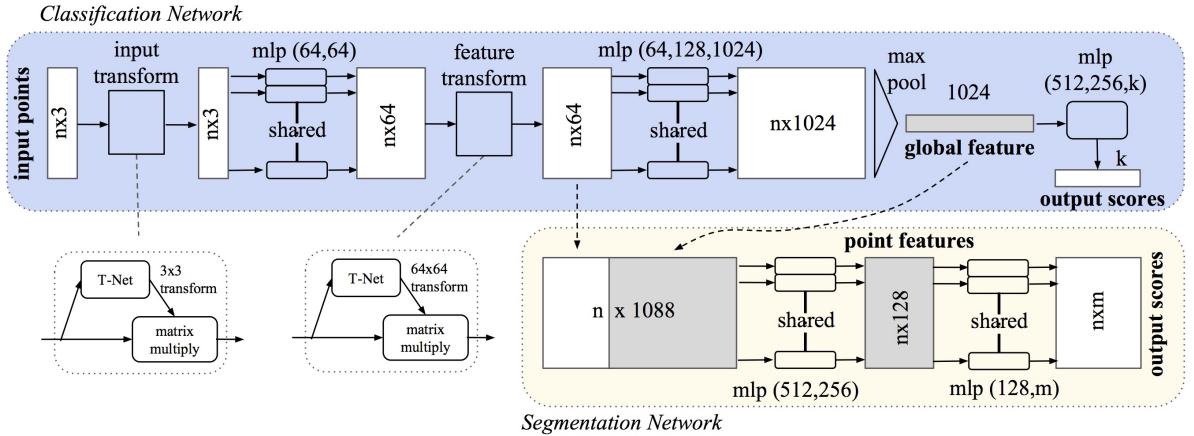


Figura 5.3: Arquitetura da PointNet. A rede de classificação tem o *input* com n ponto com 3 coordenadas, onde são aplicadas sequencias de transformação que são agregadas por uma camada de max pooling. O *output* é a classificação para k classes possíveis. A rede de segmentação é uma extensão da rede de classificação, classificando ponto a ponto a nuvem de pontos, em m classes possíveis. Mais detalhes sobre a arquitetura podem ser encontrados na Ref. [67]

A PointNet é uma rede neural capaz de processar *pointclouds*, entendendo que a estrutura dos dados é invariante à troca de pontos e invariante sobre transformações (rotação e translação)[68], pode ser usada tanto para classificação quanto segmentação semântica de nuvens de pontos[67]. A classificação é no para a *pointcloud* completa, já a segmentação semântica é a classificação ponto a ponto da nuvem de pontos[67].

Para o uso da PointNet para classificação, foi criado um banco de dados para o treino da rede neural que tem como *input* a nuvem de pontos e como *output* o número de trajetórias presentes no evento. Um dos pontos negativos da PointNet é seu tamanho fixo nos dados de entrada[67], pois no caso dos dados desse trabalho, o número de pontos por evento não é fixo. Para resolver isso, foi determinado um tamanho fixo $n = 300$ pontos para todas as nuvens de pontos. Eventos com mais pontos que n são descartados, já eventos com menos pontos eram preenchidos com pontos repetidos da mesma nuvem de pontos, para completar até o tamanho n . Eventos com menos que 100 pontos também foram descartados.

Para determinar o número de trajetórias nos eventos para o treino da rede neural, foi utilizado o estimador robusto que chamei de prototype-RANSAC (RANdom SAmple Consensus), que é uma variação do RANSAC[69, 64]. A escolha dele no lugar do HDBSCAN é pela melhor acurácia na detecção de trajetórias, apesar de ser cerca de 4 vezes mais lento. As nuvens de pontos (sem o acréscimo de pontos para a PointNet) ainda passam pelos dois filtros e critérios de correção mostrados na seção 5.1. O algoritmo

1 mostra o funcionamento do p-RANSAC. O algoritmo seleciona dois pontos de modo aleatório (*Random Sampling*) e determina o versor \hat{v} e o ponto P_b que descrevem a única reta r que passa pelos dois pontos. A reta é selecionada caso tenha um número mínimo $tam_{min} = 24$ de pontos que pertencem à reta (chamado de *inliers*) e tenha o mínimo (com relação aos outros conjunto de pontos) da estimativa C dada por[64]

$$C = \sum_{i=0}^N \frac{d_i^2}{N}, \quad (5.11)$$

onde N é o número total de pontos de uma reta e d_i é a distância do i-ésimo ponto à reta. O número de iterações do algoritmo foi determinado como sendo 700.

Algoritmo 1: p-RANSAC

Dados: pointcloud, N , d_{min} , tam_{min}

- 1 **para** cada iteração $i = 1, 2, \dots, N$ **faça**
 - 2 Seleciona dois pontos da *pointcloud* de modo aleatório (*Random Sampling*);
 - 3 Estima versor v e um ponto P_b que passe pela reta r formada pelos dois pontos;
 - 4 **para** cada ponto P **faça**
 - 5 Calcula a distância d do ponto à reta r ;
 - 6 **se** $d < d_{min}$ **então**
 - 7 Guarda P como pertencente à r ;
 - 8 **se** *Número de pontos de r* $> tam_{min}$ **então**
 - 9 Guarda v , P_b e C ;
 - 10 Ordena as retas do menor para o maior C ;
 - 11 **para** cada reta r ordenada **faça**
 - 12 **se** *Número de pontos de r* $> tam_{min}$ **então**
 - 13 Guarda v , P_b e pontos $P \in r$;
 - 14 **retorna** Retas r selecionadas na última etapa;
-

Para o *output* foram escolhidos eventos que possuem de 0 até 5 trajetórias. Com o *output* construído, a rede para classificação foi treinada, onde a função custo foi a *categorical cross entropy*[29] e o otimizador o ADAM[30]. A métrica utilizada foi a acurácia categórica e foi de 70% quando concluído o treino. Exemplo do resultado da aplicação da rede neural em nuvens de pontos está na figura 5.4.

Trajetórias = 3. Resultado PointNet = 3

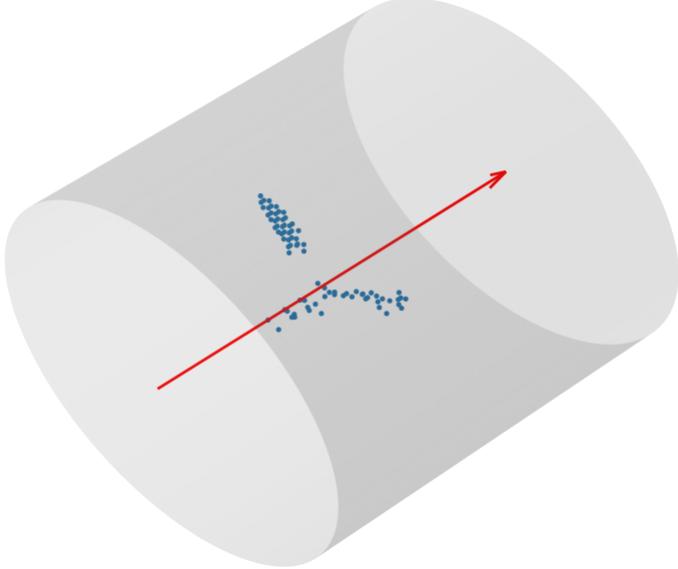


Figura 5.4: Nuvem de pontos que possui 3 trajetórias para serem detectadas e a rede neural de classificação calculou que haviam 3 trajetórias, o que indica um resultado correto do algoritmo.

Eventos com 4 ou 5 trajetórias ocorrem com uma frequência muito menor (menos que 5% do total de dados) que eventos de 1 e 2 trajetórias (cerca de 85% do total de dados), caracterizando o banco de dados com o problema de desbalanço de classe, o que pode ter prejudicado o treino para identificar eventos mais incomuns. No dados do experimento desse trabalho não foi necessário utilizar essa rede neural, pois a etapa de identificação de trajetórias é rápida em comparação com a mesma etapa usado dados gerados nos experimentos com o AT-TPC[12, 47].

5.2.2 Detecção de *outliers*

Um possível uso para a rede de segmentação semântica é para classificar pontos como *inliers* ou *outliers*, semelhante ao uso do algoritmo *outlier removal* da biblioteca Open3D[61]. A diferença é que com a PointNet pode-se incluir *outliers* locais[68], não apenas os globais para serem identificados.

Para o *output* da rede neural, os pontos classificados como *outliers* globais ou locais possuem valor 0 e pontos que são *inliers* (pertencem à alguma trajetória) possuem valor 1. A rede de segmentação foi treinada, com a função custo sendo a *binary cross entropy* e a métrica a acurácia binária. A acurácia da rede foi de aproximadamente 93%, se mostrando uma boa alternativa para eliminação de ruído nos eventos. A figura 5.5 mostra o resultado da aplicação da rede de segmentação em uma nuvem de pontos.

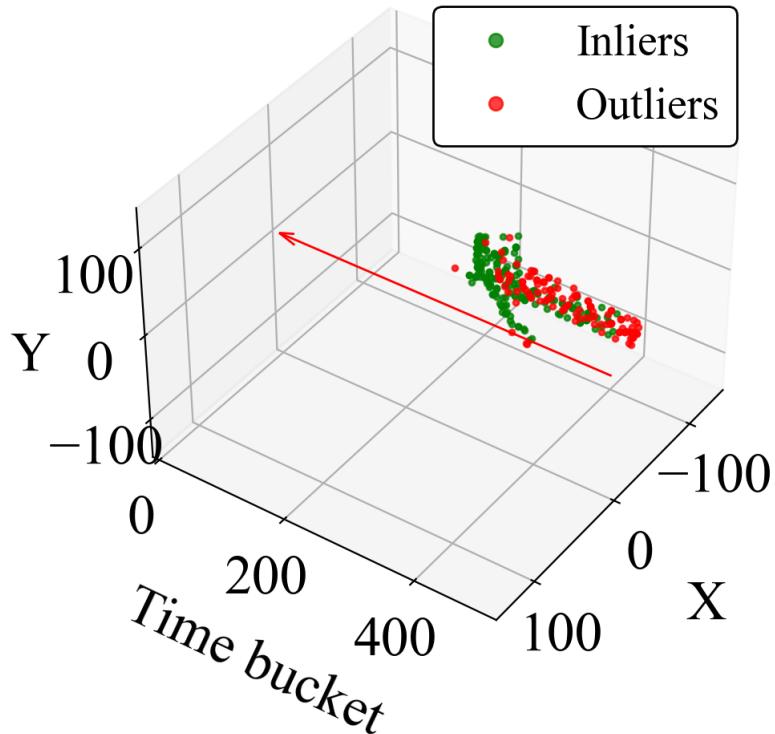


Figura 5.5: Detecção de outliers com a rede neural de segmentação. A rede neural foi capaz de detectar os *outliers* desse evento com 95% de acurácia.

Existem outras redes neurais que são capazes de lidar com *pointclouds*, como a PointNet++[70] e Dynamic Graph CNN[71]. Para a abordagem de detecção de trajetórias de forma direta, uma opção é utilizar a rede neural ContrastNet[72], pois a rede é capaz de fazer *clustering* nos dados, retornando os diferentes *clusters* existentes nos dados.

Com as informações da cinemática determinadas, foi possível determinar as distribuições angulares, que está feito no capítulo 6.

Capítulo 6

Resultados

Após identificar todos os *clusters* de cada evento devemos identificar quais são as partículas que originaram cada uma das trajetórias detectadas. Temos o comprimento de cada trajetória e sua energia, portanto o objetivo é, dada essas duas informações, identificar qual a partícula. A figura 6.1 mostra um histograma bidimensional do comprimento de cada *track* em função do ângulo de espalhamento no referencial do laboratório, usando apenas eventos que possuíam duas *tracks* com o mesmo vértice de reação. É possível perceber as medidas coincidentes do ^{16}O e do próton.

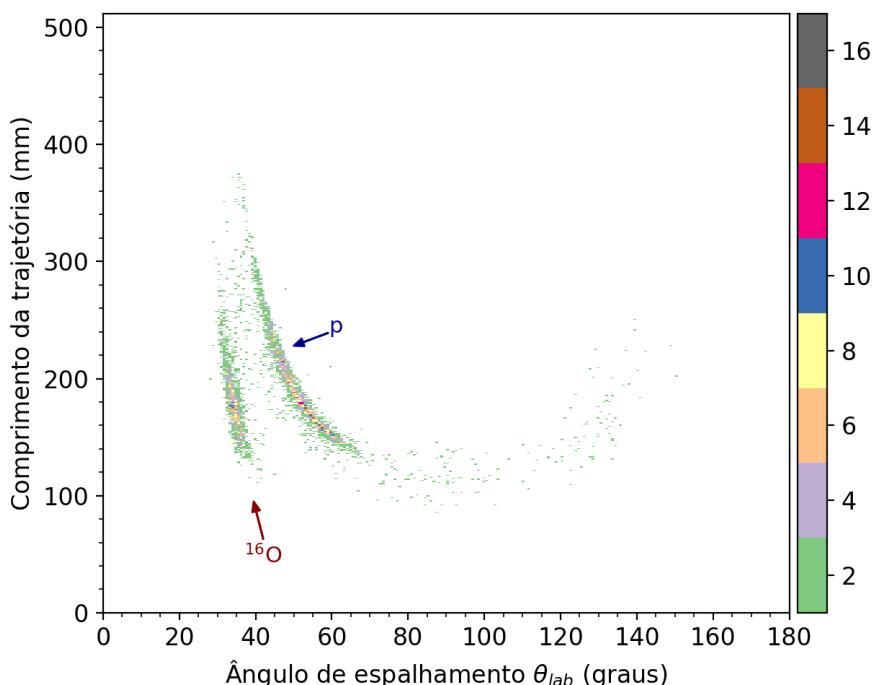


Figura 6.1: Histograma de comprimento de *track* no eixo y e ângulo de espalhamento no eixo x. O histograma foi feito coletando eventos que possuíam duas trajetórias com o mesmo vértice de reação, indicando a detecção simultânea do ^{16}O e do próton.

Para determinar qual é a partícula de cada *track* podemos usar o LISE++[73] para calcular o alcance (*range*) das possíveis partículas (^{17}F , ^{16}O e próton) dada as propriedades do alvo (^4He à uma pressão de 350 Torr). A figura ?? mostra o alcance em mm das partículas em função da energia em MeV.

Capítulo 7

Conclusão

Referências

- [1] D. Suzuki et al. “Prototype AT-TPC: Toward a new generation active target time projection chamber for radioactive beam experiments”. In: *Nuclear Instruments and Methods in Physics Research Section A Accelerators Spectrometers Detectors and Associated Equipment* 691 (Nov. 2012), p. 39. DOI: [10.1016/j.nima.2012.06.050](https://doi.org/10.1016/j.nima.2012.06.050).
- [2] F.D. Becchetti et al. “The TwinSol low-energy radioactive nuclear beam apparatus: status and recent results”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 505.1 (2003). Proceedings of the tenth Symposium on Radiation Measurements and Applications, pp. 377 –380. ISSN: 0168-9002. DOI: [https://doi.org/10.1016/S0168-9002\(03\)01101-X](https://doi.org/10.1016/S0168-9002(03)01101-X). URL: <http://www.sciencedirect.com/science/article/pii/S016890020301101X>.
- [3] J.J. Kolata et al. “A radioactive beam facility using a large superconducting solenoid”. In: *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms* 40-41 (1989), pp. 503–506. ISSN: 0168-583X. DOI: [https://doi.org/10.1016/0168-583X\(89\)91032-X](https://doi.org/10.1016/0168-583X(89)91032-X). URL: <https://www.sciencedirect.com/science/article/pii/0168583X8991032X>.
- [4] Panagiotis Gastis et al. “Improved phenomenological description of equilibrium charge state distributions for Ni, Co, and Cu ions in Mo based on new experimental data at 2 MeV/u”. In: (Sept. 2015).
- [5] L. E. Tamayose et al. “Simulation of the RIBRAS Facility with GEANT4”. In: *Brazilian Journal of Physics* 52.3 (2022), p. 89. ISSN: 1678-4448. DOI: [10.1007/s13538-022-01090-y](https://doi.org/10.1007/s13538-022-01090-y). URL: <https://doi.org/10.1007/s13538-022-01090-y>.

- [6] J.C. Zamora. “Estudo do espalhamento elástico dos isótopos 7Be , 9Be e 10Be em alvo de 12C ”. MA thesis. University of São Paulo, Brazil, 2011. DOI: 10.11606/D.43.2011.tde-30092011-132427.
- [7] S. Agostinelli et al. “Geant4—a simulation toolkit”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 506.3 (2003), pp. 250–303. ISSN: 0168-9002. DOI: [https://doi.org/10.1016/S0168-9002\(03\)01368-8](https://doi.org/10.1016/S0168-9002(03)01368-8). URL: <https://www.sciencedirect.com/science/article/pii/S0168900203013688>.
- [8] R. Lichtenhäler et al. “Radioactive ion beams in Brasil (RIBRAS)”. en. In: *Brazilian Journal of Physics* 33 (June 2003), pp. 294 –296. ISSN: 0103-9733. URL: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-97332003000200025&nrm=iso.
- [9] Sergio Muniz, Mishkat Bhattacharya, and V. Bagnato. “Simple analysis of off-axis solenoid fields using the scalar-magnetostatic potential: application to a Zeeman-slower for cold atoms”. In: (Mar. 2010).
- [10] Jaspreet Singh Randhawa et al. “Beam induced space-charge effects in Time Projection Chambers in low-energy nuclear physics experiments”. In: *Nucl. Instr. and Meth. A* (July 2019).
- [11] Y. Giomataris et al. “MICROMEGAS: a high-granularity position-sensitive gaseous detector for high particle-flux environments”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 376.1 (1996), pp. 29–35. ISSN: 0168-9002. DOI: [https://doi.org/10.1016/0168-9002\(96\)00175-1](https://doi.org/10.1016/0168-9002(96)00175-1). URL: <https://www.sciencedirect.com/science/article/pii/0168900296001751>.
- [12] J. Bradt et al. “Commissioning of the Active-Target Time Projection Chamber”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 875 (2017), pp. 65 –79. ISSN: 0168-9002. DOI: <https://doi.org/10.1016/j.nima.2017.09.013>. URL: <https://www.sciencedirect.com/science/article/pii/S0168900217309683>.
- [13] J. W. Bradt. “Measurement of isobaric analogue resonances of ^{47}Ar with the Active-Target Time Projection Chamber”. PhD thesis. Michigan State University, USA,

2017. URL: https://publications.nscl.msu.edu/thesis/%20Brandt_2017_5279.pdf.

- [14] J. Giovinazzo et al. “GET electronics samples data analysis”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 840 (2016), pp. 15–27. ISSN: 0168-9002. DOI: <https://doi.org/10.1016/j.nima.2016.09.018>. URL: <https://www.sciencedirect.com/science/article/pii/S0168900216309408>.
- [15] Thomas Lohse and Werner Witzeling. “The Time Projection Chamber”. In: *Instrumentation in High Energy Physics*, pp. 81–155. DOI: 10.1142/9789814360333_0002. eprint: https://www.worldscientific.com/doi/pdf/10.1142/9789814360333_0002. URL: https://www.worldscientific.com/doi/abs/10.1142/9789814360333_0002.
- [16] A. Geron. *Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: concepts, tools, and techniques to build intelligent systems*. 2nd ed. O’Reilly, 2019. ISBN: 9781492032649.
- [17] *Machine Learning and Data Analysis for Nuclear Physics*. 2020. URL: <https://ectstar.fbk.eu/node/4472>.
- [18] Maryam M. Najafabadi et al. “Deep learning applications and challenges in big data analytics”. In: *Journal of Big Data* 2.1 (2015), p. 1. ISSN: 2196-1115. DOI: 10.1186/s40537-014-0007-7. URL: <https://doi.org/10.1186/s40537-014-0007-7>.
- [19] Morten Hjorth-Jensen. *Data Analysis and Machine Learning: Neural networks, from the simple perceptron to deep learning*. 2020. URL: <https://nucleartalent.github.io/MachineLearningECT/doc/pub/Day4/pdf/Day4.pdf> (visited on 08/2020).
- [20] C.M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer New York, 2016. ISBN: 9781493938438. URL: <https://books.google.com.br/books?id=kOXDtAEACAAJ>.
- [21] Steven B. Damelin and Willard Miller Jr. *The Mathematics of Signal Processing*. Cambridge Texts in Applied Mathematics. Cambridge University Press, 2011. DOI: 10.1017/CBO9781139003896.

- [22] K.-L Du and M.N.s Swamy. “Radial Basis Function Networks”. In: Dec. 2014, pp. 299–335. ISBN: 978-1-4471-5570-6. DOI: 10.1007/978-1-4471-5571-3_10.
- [23] Alex Sherstinsky. “Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network”. In: *Physica D: Nonlinear Phenomena* 404 (2020), p. 132306. ISSN: 0167-2789. DOI: <https://doi.org/10.1016/j.physd.2019.132306>. URL: <https://www.sciencedirect.com/science/article/pii/S0167278919305974>.
- [24] Abien Fred Agarap. “Deep Learning using Rectified Linear Units (ReLU)”. In: (Mar. 2018).
- [25] Moraga C. Han J. “The influence of the sigmoid function parameters on the speed of backpropagation learning”. In: *Lecture Notes in Computer Science* 930 (1995). DOI: https://doi.org/10.1007/3-540-59497-3_175.
- [26] Tomasz Szandała. *Review and Comparison of Commonly Used Activation Functions for Deep Neural Networks*. Oct. 2020. DOI: https://doi.org/10.1007/978-981-15-5495-7_11.
- [27] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. <http://www.deeplearningbook.org>. MIT Press, 2016.
- [28] Hecht-Nielsen. “Theory of the backpropagation neural network”. In: *International 1989 Joint Conference on Neural Networks*. 1989, 593–605 vol.1. DOI: 10.1109/IJCNN.1989.118638.
- [29] Douglas Kline and Victor Berardi. “Revisiting squared-error and cross-entropy functions for training neural network classifiers”. In: *Neural Computing and Applications* 14 (Dec. 2005), pp. 310–318. DOI: 10.1007/s00521-005-0467-y.
- [30] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *CoRR* abs/1412.6980 (2015).
- [31] Sebastian Ruder. “An overview of gradient descent optimization algorithms”. In: *ArXiv* abs/1609.04747 (2016).
- [32] Mohammad Hossin and Sulaiman M.N. “A Review on Evaluation Metrics for Data Classification Evaluations”. In: *International Journal of Data Mining & Knowledge Management Process* 5 (Mar. 2015), pp. 01–11. DOI: 10.5121/ijdkp.2015.5201.

- [33] N. S. Altman. “An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression”. In: *The American Statistician* 46.3 (1992), pp. 175–185. DOI: 10.1080/00031305.1992.10475879. eprint: <https://www.tandfonline.com/doi/pdf/10.1080/00031305.1992.10475879>. URL: <https://www.tandfonline.com/doi/abs/10.1080/00031305.1992.10475879>.
- [34] Angela Serra and Roberto Tagliaferri. “Unsupervised Learning: Clustering”. In: Jan. 2018. ISBN: 9780128096338. DOI: 10.1016/B978-0-12-809633-8.20487-1.
- [35] Kristina Sinaga and Miin-Shen Yang. “Unsupervised K-Means Clustering Algorithm”. In: *IEEE Access* PP (Apr. 2020), pp. 1–1. DOI: 10.1109/ACCESS.2020.2988796.
- [36] Y Reddy, Viswanath Pulabaigari, and Eswara B. “Semi-supervised learning: a brief review”. In: *International Journal of Engineering & Technology* 7 (Feb. 2018), p. 81. DOI: 10.14419/ijet.v7i1.8.9977.
- [37] Michael L Littman Leslie Pack Kaelbling and Andrew W Moore. “Reinforcement learning: A survey”. In: *Journal of artificial intelligence research* (1996), 237–285. DOI: <https://doi.org/10.1613/jair.301>.
- [38] Kim D et al. “Review of machine learning methods in soft robotics”. In: *PLoS ONE* (2021). DOI: <https://doi.org/10.1371/journal.pone.0246102>.
- [39] Stefano Carboni et al. “Particle identification using the (DELT)A-E-E technique and pulse shape discrimination with the silicon detectors of the FAZIA project”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 664 (Feb. 2012), 251–263. DOI: 10.1016/j.nima.2011.10.061.
- [40] Erich Schubert et al. “DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN”. In: *ACM Trans. Database Syst.* 42.3 (2017). ISSN: 0362-5915. DOI: 10.1145/3068335. URL: <https://doi.org/10.1145/3068335>.
- [41] Guolin Ke et al. “LightGBM: A Highly Efficient Gradient Boosting Decision Tree”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS’17. Long Beach, California, USA: Curran Associates Inc., 2017, 3149–3157. ISBN: 9781510860964.

- [42] Ze-Peng Gao et al. “Machine learning the nuclear mass”. In: *Nuclear Science and Techniques* 32.10 (2021), p. 109. ISSN: 2210-3147. DOI: 10.1007/s41365-021-00956-1. URL: <https://doi.org/10.1007/s41365-021-00956-1>.
- [43] R Utama, Wei-Chia Chen, and J Piekarewicz. “Nuclear charge radii: density functional theory meets Bayesian neural networks”. In: *Journal of Physics G: Nuclear and Particle Physics* 43.11 (2016), p. 114002. DOI: 10.1088/0954-3899/43/11/114002. URL: <https://doi.org/10.1088/0954-3899/43/11/114002>.
- [44] Niu Zhongming et al. “Predictions of nuclear β -decay half-lives with machine learning and their impact on r-process nucleosynthesis”. In: *Physical Review C* 99 (June 2019). DOI: 10.1103/PhysRevC.99.064307.
- [45] M.P. Kuchera et al. “Machine learning methods for track classification in the AT-TPC”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 940 (2019), pp. 156–167. ISSN: 0168-9002. DOI: <https://doi.org/10.1016/j.nima.2019.05.097>. URL: <https://www.sciencedirect.com/science/article/pii/S0168900219308046>.
- [46] Christoph Dalitz, Tilman Schramke, and Manuel Jeltsch. “Iterative Hough Transform for Line Detection in 3D Point Clouds”. In: *Image Processing On Line* 7 (2017). <https://doi.org/10.5201/ipol.2017.208>, pp. 184–196.
- [47] G.F. Fortino et al. “Digital signal analysis based on convolutional neural networks for active target time projection chambers”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 1031 (2022), p. 166497. ISSN: 0168-9002. DOI: <https://doi.org/10.1016/j.nima.2022.166497>. URL: <https://www.sciencedirect.com/science/article/pii/S016890022200119X>.
- [48] P. Holl et al. “Deep learning based pulse shape discrimination for germanium detectors”. In: *The European Physical Journal C* 79.6 (2019), p. 450. ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-019-6869-2. URL: <https://doi.org/10.1140/epjc/s10052-019-6869-2>.
- [49] Edmundo Capelas de Oliveira and José Emilío Maiorino. *Introdução aos métodos da matemática aplicada*. 3rd ed. Editora UNICAMP, 2010. ISBN: 9788526809062.
- [50] *ROOT Data Analysis Framework*. 2021. URL: <https://root.cern.ch/>.

- [51] C.G. Ryan et al. “SNIP, a statistics-sensitive background treatment for the quantitative analysis of PIXE spectra in geoscience applications”. In: *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms* 34.3 (1988), pp. 396–402. ISSN: 0168-583X. DOI: [https://doi.org/10.1016/0168-583X\(88\)90063-8](https://doi.org/10.1016/0168-583X(88)90063-8). URL: <https://www.sciencedirect.com/science/article/pii/0168583X88900638>.
- [52] Miroslav Morháč et al. “Background elimination methods for multidimensional coincidence γ -ray spectra”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 401.1 (1997), pp. 113–132. ISSN: 0168-9002. DOI: [https://doi.org/10.1016/S0168-9002\(97\)01023-1](https://doi.org/10.1016/S0168-9002(97)01023-1). URL: <https://www.sciencedirect.com/science/article/pii/S0168900297010231>.
- [53] Donald D. Burgess and Richard J. Tervo. “Background estimation for gamma-ray spectrometry”. In: *Nuclear Instruments and Methods in Physics Research* 214.2 (1983), pp. 431–434. ISSN: 0167-5087. DOI: [https://doi.org/10.1016/0167-5087\(83\)90612-9](https://doi.org/10.1016/0167-5087(83)90612-9). URL: <https://www.sciencedirect.com/science/article/pii/0167508783906129>.
- [54] Miroslav Morháč et al. “Identification of peaks in multidimensional coincidence γ -ray spectra”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 443.1 (2000), pp. 108–125. ISSN: 0168-9002. DOI: [https://doi.org/10.1016/S0168-9002\(99\)01005-0](https://doi.org/10.1016/S0168-9002(99)01005-0). URL: <https://www.sciencedirect.com/science/article/pii/S0168900299010050>.
- [55] Pauli Virtanen et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. In: *Nature Methods* 17 (2020), pp. 261–272. DOI: [10.1038/s41592-019-0686-2](https://doi.org/10.1038/s41592-019-0686-2).
- [56] Matt Newville et al. *lmfit/lmfit-py 1.0.2*. Version 1.0.2. Feb. 2021. DOI: [10.5281/zenodo.4516651](https://doi.org/10.5281/zenodo.4516651). URL: <https://doi.org/10.5281/zenodo.4516651>.
- [57] Xavier Glorot and Y. Bengio. “Understanding the difficulty of training deep feed-forward neural networks”. In: *Journal of Machine Learning Research - Proceedings Track 9* (Jan. 2010), pp. 249–256.

- [58] Ekaba Bisong. “Google Colaboratory”. In: *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*. Berkeley, CA: Apress, 2019, pp. 59–64. ISBN: 978-1-4842-4470-8. DOI: 10.1007/978-1-4842-4470-8_7. URL: https://doi.org/10.1007/978-1-4842-4470-8_7.
- [59] Andrea Dal Pozzolo et al. “Calibrating Probability with Undersampling for Unbalanced Classification”. In: Dec. 2015. DOI: 10.1109/SSCI.2015.33.
- [60] Ashraf A Aly, Safaai Bin Deris, and Nazar Zaki. “Research review for digital image segmentation techniques”. In: *Int. J. Comput. Sci. Inf. Technol. Res.* 3.5 (2011), p. 99.
- [61] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. “Open3D: A Modern Library for 3D Data Processing”. In: *arXiv:1801.09847* (2018).
- [62] Leland McInnes, John Healy, and Steve Astels. “hdbscan: Hierarchical density based clustering”. In: *The Journal of Open Source Software* 2.11 (2017), p. 205.
- [63] Leland McInnes and John Healy. “Accelerated Hierarchical Density Based Clustering”. In: *Data Mining Workshops (ICDMW), 2017 IEEE International Conference on*. IEEE. 2017, pp. 33–42.
- [64] J.C. Zamora and G.F. Fortino. “Tracking algorithms for TPCs using consensus-based robust estimators”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* (2020), p. 164899. ISSN: 0168-9002. DOI: <https://doi.org/10.1016/j.nima.2020.164899>. URL: <http://www.sciencedirect.com/science/article/pii/S0168900220312961>.
- [65] Sam Fletcher and Md Islam. “Comparing sets of patterns with the Jaccard index”. In: *Australasian Journal of Information Systems* 22 (Mar. 2018). DOI: 10.3127/ajis.v22i0.1538.
- [66] Peter J. Rousseeuw. “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis”. In: *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53 –65. ISSN: 0377-0427. DOI: [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7). URL: <http://www.sciencedirect.com/science/article/pii/0377042787901257>.
- [67] Charles R Qi et al. “PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation”. In: *arXiv preprint arXiv:1612.00593* (2016).

- [68] JMaocheng Tang. “Semantic Classification of 3D Point Cloud by Random Forest-based Feature Learning”. MA thesis. Technische Universität Berlin, Germany, 2019.
- [69] Yassid Ayyad et al. “Novel particle tracking algorithm based on the Random Sample Consensus Model for the Active Target Time Projection Chamber (AT-TPC)”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 880 (2018), pp. 166 – 173. ISSN: 0168-9002. DOI: <https://doi.org/10.1016/j.nima.2017.10.090>. URL: <http://www.sciencedirect.com/science/article/pii/S0168900217311798>.
- [70] Charles R Qi et al. “PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space”. In: *arXiv preprint arXiv:1706.02413* (2017).
- [71] Yue Wang et al. *Dynamic Graph CNN for Learning on Point Clouds*. 2018. DOI: 10.48550/ARXIV.1801.07829. URL: <https://arxiv.org/abs/1801.07829>.
- [72] Ling Zhang and Zhigang Zhu. *Unsupervised Feature Learning for Point Cloud by Contrasting and Clustering With Graph Convolutional Neural Network*. 2019. DOI: 10.48550/ARXIV.1904.12359. URL: <https://arxiv.org/abs/1904.12359>.
- [73] MP Kuchera et al. “LISE++ Software Updates and Future Plans”. In: *Journal of Physics: Conference Series* 664.7 (2015), p. 072029. DOI: 10.1088/1742-6596/664/7/072029. URL: <https://doi.org/10.1088/1742-6596/664/7/072029>.