Machine Learning (ML) for Improved Analyses of High Resolution Gaseous Detector Data

Dr. Jeff Schueler*
CPAD 2023: Nov 9th, 2023

*Email: schuel93@gmail.com
Detector technology highlighted today

**BEAST TPCs: 40 cm³ pixel readout CYGNUS prototype directional detection vessels**

- 70:30 He:CO₂ 1 atm
- Double thin GEM amplification
- 2cm x 1.68cm pixel readout with 250μm x 50μm segmentation

**Goal:** Demonstrate directionality at the lowest possible NR energies

**Pixel-level 3D reconstruction**

~300 keV He recoil

**ATLAS FE-I4b**

---

**MIGDAL TPC: 108 cm³ with CMOS camera, strip, and PMT**

- 50 torr CF₄
- Double glass GEM
- High dynamic range
- 8cm x 4.5cm optical readout, 39μm seg.

**Goal:** Directly measure the Migdal effect in nuclear scattering

---

*This talk only focuses on the camera readout!

---

The ML techniques described in these slides are not restricted to pixel and optical readouts!
Many implementations of algorithms for object detection and key point detection already exist (see here and here), and there are many tools for developing custom classification and regression models for 2D and 3D image data (PyTorch tutorial, TensorFlow tutorial, spconv for sparse 3D convolutional neural networks).

Today I’ll highlight examples of (1) directional reconstruction for a CYGNUS prototype BEAST TPC and (2) object detection for the rare event Migdal search.
Deep learning to assign principal axis and head/tail direction (3D reconstruction)

**Model**

**Custom ResNet 50** *(Architecture in backup)*

Train model to minimize total loss, $L$ (encodes the difference between prediction and truth)

**GOOD CHOICE OF LOSS FUNCTION IS ESSENTIAL!**

$$L \equiv L_1(y_{\text{pred},1}, y_{\text{truth},1}) + \alpha L_2(p_{\text{pred},2}, y_{\text{truth},2})$$

- We set $\alpha$ to 1
- $L_1 \equiv 1 - |\cos(y_{\text{pred},1}, y_{\text{truth},1})|$ *(Ang. resolution loss)*
- $L_2 \equiv -(y_{\text{truth},2} \log(p_{\text{pred},2}) + (1-y_{\text{truth},2}) \log(1-p_{\text{pred},2}))$ *(Head/tail loss)*

Training Data

$x$

**Labels**

$y_{\text{truth},1}$, $y_{\text{truth},2}$

Output 1:
Recoil axis

- $y_{\text{pred},1}$ *(1 x 3) axis*
- $L_1(y_{\text{pred},1}, y_{\text{truth},1})$

Output 2:
Head/tail assignment

- $y_{\text{truth},1}$
- Truth axis
- $p_{\text{pred},2}$
- Probability
- $L_2(p_{\text{pred},2}, y_{\text{truth},2})$
- $y_{\text{truth},2}$
- 0 or 1
Testing our model on head/tail recognition

~ 145 hours of data

Assign true direction as -x

~ 185 hours of data

Assign true direction as +x

~ 340 hours of data

ResNet is trained on simulation, we evaluate its head/tail recognition on measured data. We don’t have angular resolution measurements to test on*

*angular resolution results on simulation in backup
Results on He recoils

**Measurement, 50% He-recoil efficiency**

ResNet yields significant improvement in $\varepsilon_{ht}$ on measurement → First significant $\varepsilon_{ht}$ below 20 keV<sub>ee</sub> in an atmospheric pressure gas mixture!
Rare event searches: The Migdal effect

- Migdal effect is rare and has never been observed in nuclear scattering
  - Migdal signal with 5-15 keV ER expected in ~1 in 50,000 nuclear recoils ($E > 100$ keV) induced by 2.5 MeV neutrons in 50 torr CF$_4$ ([details here](#))

See E. Tilly’s talk for a more thorough intro to the MIGDAL experiment

The MIGDAL experiment is designed to record (2,048 x 1,152) images at a rate of up to 120 Hz
→ Over 1 PB of data per month!
→ Analyzing this amount of data in a timely manner is challenging, especially in a small experiment, but machine learning can help us!
Enter YOLO (You only look once)

YOLOv8 is a state of the art object detection algorithm that simultaneously locates (draws a bounding box) and identifies objects of interest in an image.

We train YOLOv8 on measured data to identify ERs, NRs, protons, alphas, sparks, camera afterglow, rolling shutter, etc.

**Benefits:**
1. Can identify multiple particle species within a continuous cluster.
2. Not trained specifically to find Migdal candidates → robust and doesn’t need to be trained on simulation!
4. Enables real time $^{55}$Fe calibrations and ER/NR event rate counting.
The data processing pipeline

(1) Dark subtracted image

(2) Downsample

**At this stage we train YOLOv8 by hand labeling bounding boxes**

(3) YOLOv8 predicts bounding boxes

Retrain as needed

(4) Perform analysis on each bounding box, computing qtys such as: Intensity, track length, angle (with head/tail), bounding box centroid

(5) Save coordinates of each bounding box, as well as extracted physics information

This entire pipeline runs at 200 fps on a consumer desktop GPU, reduces data size by a factor of ~5,000 and is integrated with the MIGDAL DAQ → Real time feedback!

CPAD Workshop 2023
Application to MIGDAL searches

- Initial science run recorded from July 17th 2023 - August 3rd 2023
- Collected an unblinded dataset consisting of 10 million 2,048 x 1,152 images
  - 20ms exposure → We expect coincidences in the camera frames

1. Initial sample of (10 million frames)
2. Find frames containing at least 1 ER and 1 NR (~25,000 frames)
3. Map bounding boxes to raw image and compute the distances between NR vertices and centers of ER bounding boxes
4. Filter frames based on distance between ER and NR (~1,000 frames with $d < 7.5\text{mm}$ and NR > 100 keV$_{\text{ee}}$)
Migdal identification performance

- Evaluate YOLOv8 (trained on real data) on simulation to determine a Migdal detection efficiency
  - **Background sample:** ~50,000 frames containing an ER and NR with random uniform separation
  - **Signal sample:** ~2,000 frames with simulated Migdals that contain ER pixels outside of the NR

- We expect the 64% Migdal efficiency evaluated on simulation to be a lower limit of performance (it should do better when applied to real data)

**YOLOv8 bounding box analysis takes us from 10’s of millions of frames to a few thousand frames while maintaining at least 64% of the Migdal signal → No longer a rare event search! We can spend the rest of our resources optimizing signal purity!**
Summary

- ResNets improve low energy directional reconstruction in a pixel-readout TPC
  - He-recoil head/tail sensitivities below 20 keV$_{ee}$ in 1atm He:CO$_2$
- YOLOv8 enables real time analysis of high throughput, high resolution image data
  - Able to analyze 100TB of data per day on a consumer desktop PC with a single graphics card
  - Reduction from 10's of millions to thousands of image frames turns a rare event search into an ordinary event search, opening up many avenues for signal purity optimization!

Modern deep learning approaches are more accessible and user friendly than ever for improving analyses of gas detector data!
Backup
Quantifying 3D directional reconstruction (BEAST TPCs)

Angular resolution: The mean difference in 3D angle, $\theta$, between the true recoil axis, $v_1$, and the measured recoil axis, $v_2$.

- Perfect angular resolution: $0^\circ$
- No angular resolution: $57^\circ$

Head/tail recognition efficiency ($\epsilon_{ht}$): Fraction of events where the dot product of the reconstructed vector and the true initial recoil direction is positive

$$\epsilon_{ht} = 0.5 \iff \text{No head/tail sensitivity}$$
$$\epsilon_{ht} = 1 \iff \text{Perfect head/tail sensitivity}$$

We train a custom 50 layer ResNet, to predict (1) a recoil axis (angular resolution), and (2) a vector (head/tail) direction. To train it, we assign two labels to each piece of training data:

Label 1 (Angular resolution): Truth recoil axis
Label 2 (Head/tail): 1 if track points in +x direction, 0 if track points in –x direction
Data vs simulation in BEAST TPCs

Simulation schematic

- Primary track
- Diffusion
- Amplification
- Digitization
3D angular and directional reconstruction in the ResNet 50 architecture
Noticeable improvement in angular resolution evaluated on simulation. We don’t have measurements to test angular resolution on data, but we can test head/tail on data!
Expected Migdal backgrounds per 1 million DD-induced nuclear recoils with $E > 100$ keV

<table>
<thead>
<tr>
<th>Component</th>
<th>Topology</th>
<th>D–D neutrons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$&gt;0.5$</td>
</tr>
<tr>
<td>Recoil-induced $\delta$-rays</td>
<td>Delta electron from NR track origin</td>
<td>$\approx 0$</td>
</tr>
<tr>
<td>Particle-Induced X-ray Emission (PIXE)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X-ray emission</td>
<td>Photoelectron near NR track origin</td>
<td>$1.8$</td>
</tr>
<tr>
<td>Auger electrons</td>
<td>Auger electron from NR track origin</td>
<td>$19.6$</td>
</tr>
<tr>
<td>Bremsstrahlung processes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quasi-Free Electron Br. (QFEB)</td>
<td>Photoelectron near NR track origin</td>
<td>$112$</td>
</tr>
<tr>
<td>Secondary Electron Br. (SEB)</td>
<td>Photoelectron near NR track origin</td>
<td>$115$</td>
</tr>
<tr>
<td>Atomic Br. (AB)</td>
<td>Photoelectron near NR track origin</td>
<td>$70$</td>
</tr>
<tr>
<td>Nuclear Br. (NB)</td>
<td>Photoelectron near NR track origin</td>
<td>$\approx 0$</td>
</tr>
<tr>
<td>Neutron inelastic $\gamma$-rays</td>
<td>Compton electron near NR track origin</td>
<td>$1.6$</td>
</tr>
<tr>
<td>Random track coincidences</td>
<td></td>
<td></td>
</tr>
<tr>
<td>External $\gamma$- and X-rays</td>
<td>Photo-/Compton electron near NR track</td>
<td>$\approx 0$</td>
</tr>
<tr>
<td>Trace radioisotopes (gas)</td>
<td>Electron from decay near NR track origin</td>
<td>$0.2$</td>
</tr>
<tr>
<td>Neutron activation (gas)</td>
<td>Electron from decay near NR track origin</td>
<td>0</td>
</tr>
<tr>
<td>Muon-induced $\delta$-rays</td>
<td>Delta electron near NR track origin</td>
<td>$\approx 0$</td>
</tr>
<tr>
<td>Secondary nuclear recoil fork</td>
<td>NR track fork near track origin</td>
<td>–</td>
</tr>
<tr>
<td><strong>Total background</strong></td>
<td>Sum of the above components</td>
<td>1.5</td>
</tr>
<tr>
<td><strong>Migdal signal</strong></td>
<td>Migdal electron from NR track origin</td>
<td>32.6</td>
</tr>
</tbody>
</table>
Fe55 spectra with YOLO

MIG_Fe55_600V_230803T104313.CAL

- Fe55 data
- $\mu = 342.569 \pm 1.22$
- $\sigma = 88.15$
- $\chi^2$/dof = 1.74
- Outside of fit boundary
Real time energy vs length output from the object detection pipeline