Classifying Scatters: Single or Multiple

Applying machine learning to the search for dark matter

Lucas Fenaux & Georges Kanaan – May to August 2020

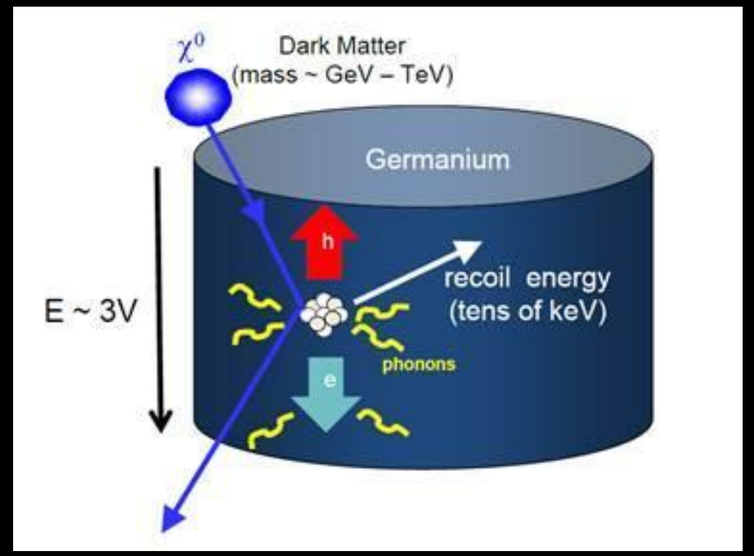
The Goal

- Use Machine Learning to help classify particle hits
- Specifically, identify whether a particle collided with the detector once or multiple times
- By achieving the above, allow for better detector calibration and further particle discrimination in the search for dark matter

Prerequisite Knowledge

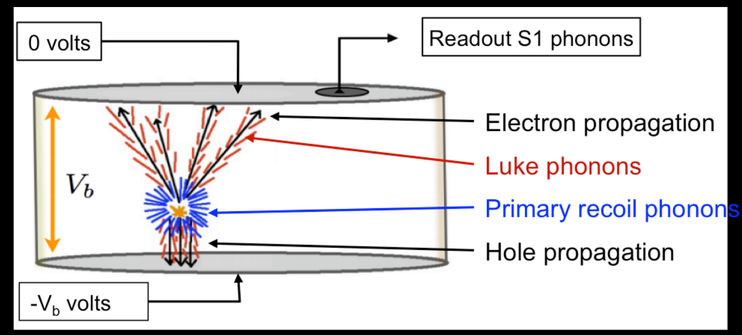
What is the SCDMS experiment?

- Dark matter may be made of elementary particles that interact only weakly with the normal matter described by the Standard Model of particle physics.
- The Cryogenic Dark Matter Search (SuperCDMS) is one of several collaborations performing experiments to detect these particles and understand the nature of the dark matter.



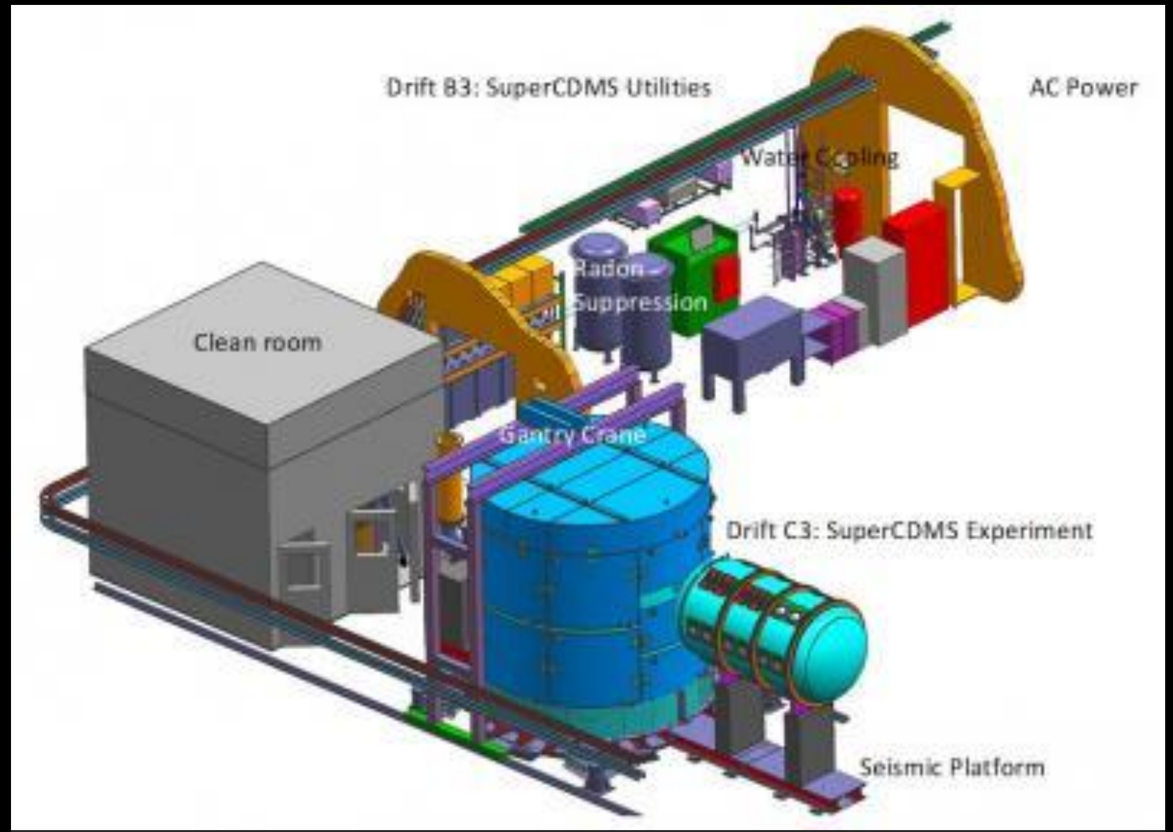
How does a particle interact with a detector?

- WIMPs and neutrons collide with the detector, scattering from atomic nuclei and depositing energy. They can interact one or multiple times with the detector.
- Electric charges and primary recoil phonons from the interactions yield a voltage readout.



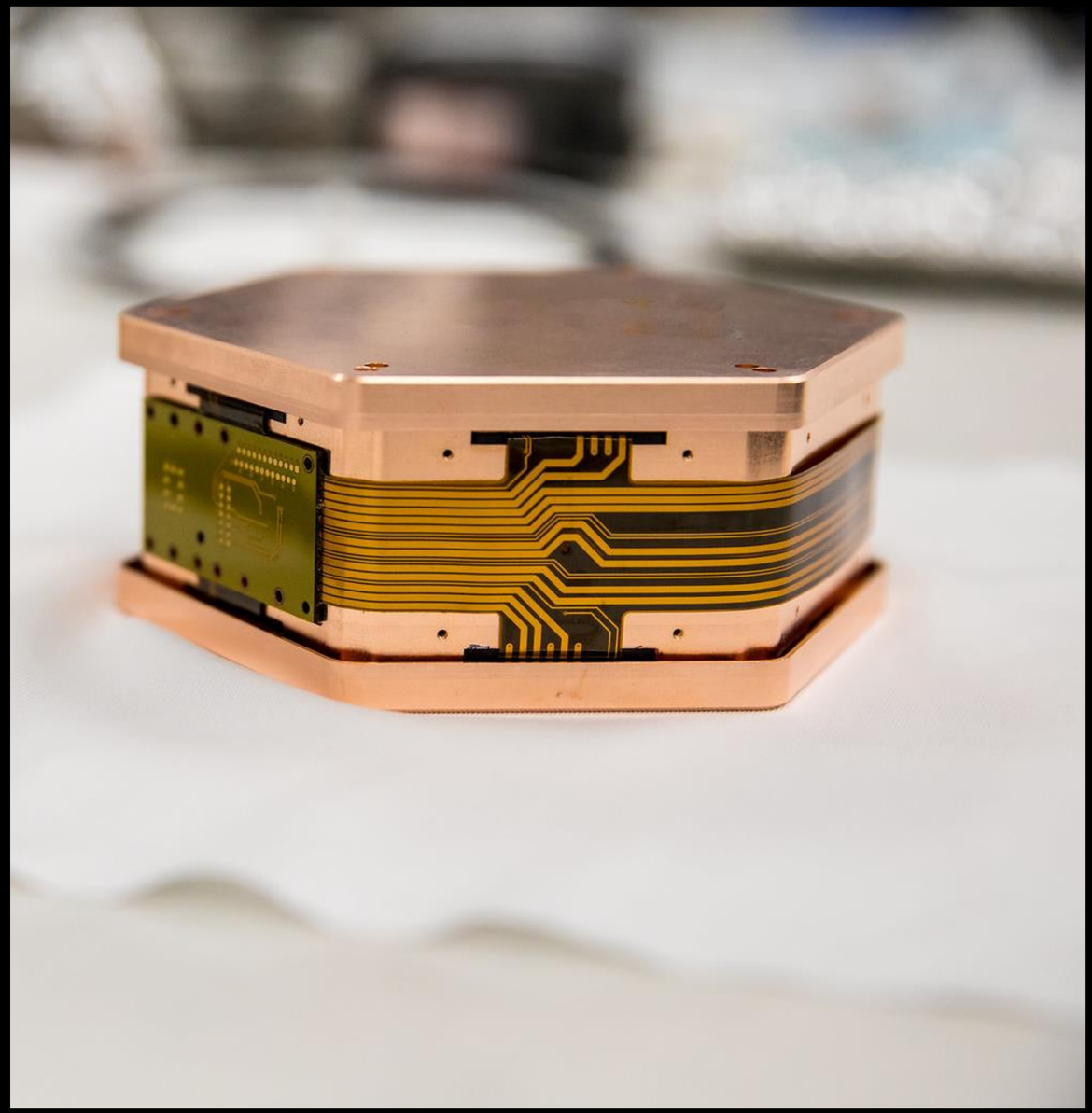
Experimental Setup

- SuperCDMS@SNOLAB is the successor to the previous generation of CDMS experiments which were in Soudan, MN
- Three detectors are placed vertically
- We are given labelled data from simulations of this process
- We are given real experimental data
- We can calculate interesting features from the data (RQs & RRQs)



The Challenge

- Various unknowns that are hard to estimate
- No available heuristic guiding the discrimination
- Reference data is only given by simulations
- Available data is limited
- No labelled real experiment data

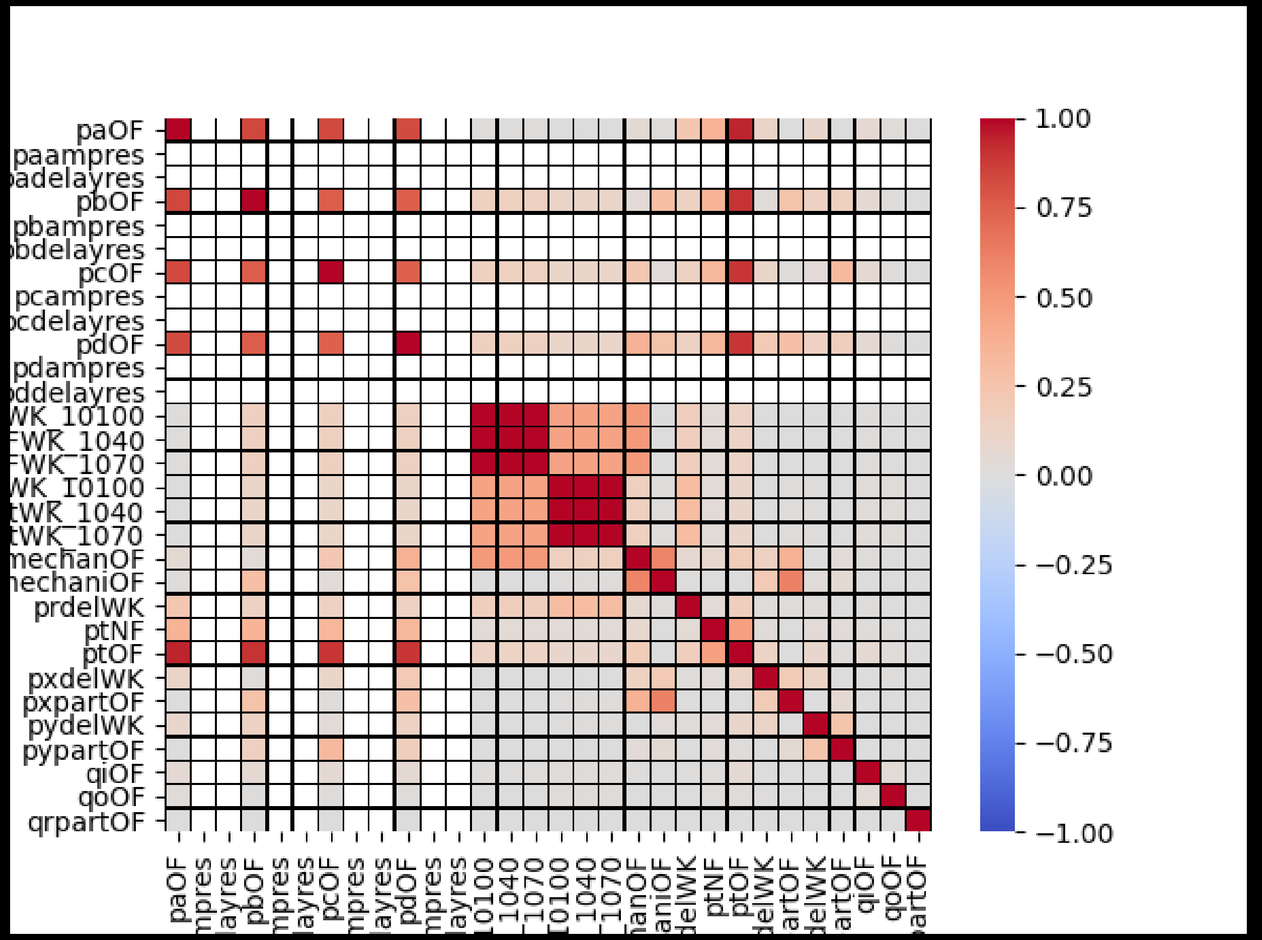


Our Approach

- We sought to build a classifier on labelled (simulated) data first.
- We use feature generating processes such as PCA to get a better visual understanding of the data we were working with. We also used feature selecting processes to isolate the most relevant parts of the features. We then pass the result through various supervised and unsupervised classifiers such a feed forward network and a k-means clustering model.
- Once we have validated a classifier on simulated, we apply it to real data.

Understanding RQ & RRQs

- We want to understand the data we're giving the model.
- We identify the most relevant RQs and RRQs
- We plot a correlation matrix to asses the robustness of the features selected
- Each RRQ's name encodes what value it represents, for example:
  - pxpartOF: phonon x partition optimal filter
  - pxdelWK: phonon x delay



Experiments & Results

- We also attempted more complex clustering methods like OPTICS to relabel the data (and also label the real data) and then training a classifier on that data
- For the other non Neural-Net methods, we tried BDTs (66.9% accuracy), Random Forests (73.3% accuracy) and Gradient Boosting (73.4% accuracy)
- Best results achieved by a FeedForward neural net and the Gradient Boosting method. Promising results are achieved through OPTICS + FeedForward.
- We can see a local minima that's reached by both the FeedForward neural net, the Gradient Boosting method and the Random Forests

	Simulated	**Real***	**Simulated + Real***
FeedForward	73.4%	NA	NA
K-Means	65%	66%	63%
OPTICS + FeedForward	71.5%	NA	69.5%
Other non-NN methods	73.4%	NA%	NA%

* real data is taken from the Soudan photroneutron dataset

Future Work

- Leverage the nature of raw data to give the model an edge
- Use LSTMs in combination with more advanced models on the raw data
- Refine the feature extraction and feature generation methods to hone and augment the data.
- Acquire more data to give more complex structures like OPTICS + FeedForward a better chance

Thank You