Simulations and Machine Learning in Dark Matter Searches

Classification methods, optimal filtering, best estimates extraction

Jorge D. Morales Mendoza

03/23/2022
Seminarios DATAI
Universidad de Navarra

Outline

I. Motivation: Dark Matter (DM) and Existence of WIMPs

A. Quick overview of the DM experiments
B. The DM search challenge

II. The Use of Simulations and ML in DM Search Experiments

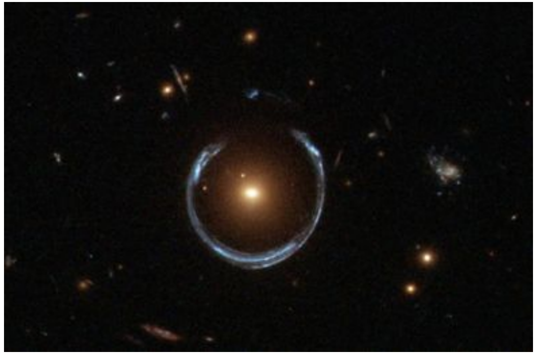
A. Automated processing pipeline
B. Simulations pipeline
C. A needle in a haystack: classification methods
D. Assembling the haystack: log-likelihood fitting of the background model
E. Best estimates of *physical* quantities from raw electronic signals: optimal filtering and more

III. Conclusions

I. Motivation: Dark Matter (DM) and the Existence of WIMPs

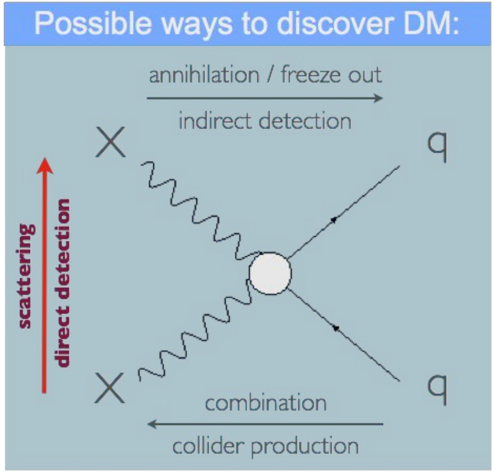
Motivation: Dark Matter (DM) and Existence of WIMPs

- Dark Matter is 85% of the gravitational mass of the Universe
  → never directly measured!
- Summary of evidence for DM:
  - Rotational velocity of stars in outer parts of galaxies suggests greater than visible (light-interacting) mass. DM is distributed in a large halo filling each galaxy, including our own Milky Way
  - Gravitational bend (Einstein ring) suggests greater than observed mass
  - Cosmic Background Radiation measurements indicate that there is a large amount of mass in the universe not in atoms (or SM particles)
  - Colliding clusters of galaxies provide evidence that DM is likely to be a particle
- Simplest guess is that DM is a WIMP:
  - Weakly interacting, Massive Particle
  - Neutral, but neutrinos are ruled out
  - Most believe it must be a new type of particle (example theories are Supersymmetry and Dark Sector)



Possible Ways of Detecting WIMPs

- Looking at the non-gravitational WIMP interaction/coupling with the Standard Model
  - Indirect Detection:
    Astronomical observations from WIMP annihilation which will produce "anomalous" high energy SM particles
  - Collider Production:
    High energy collisions of SM particles, producing WIMPs
  - **Direct Detection**:
    Earth is expected to be immersed in the Dark Matter of the Milky Way so it should be flowing through the detectors and hopefully we can detect an interaction. *This work is about searching for WIMPs this way*



- **Current state of the art:**
  - WIMPs haven't been found by any experiment, but many are looking
  - Limits on the likelihood of interaction and its dependence on the WIMP mass keep pushing to better sensitivities
  - *LZ* is (soon will be!) world leading at searches in a *broad* WIMP mass spectrum (10 to 10^4 GeV)

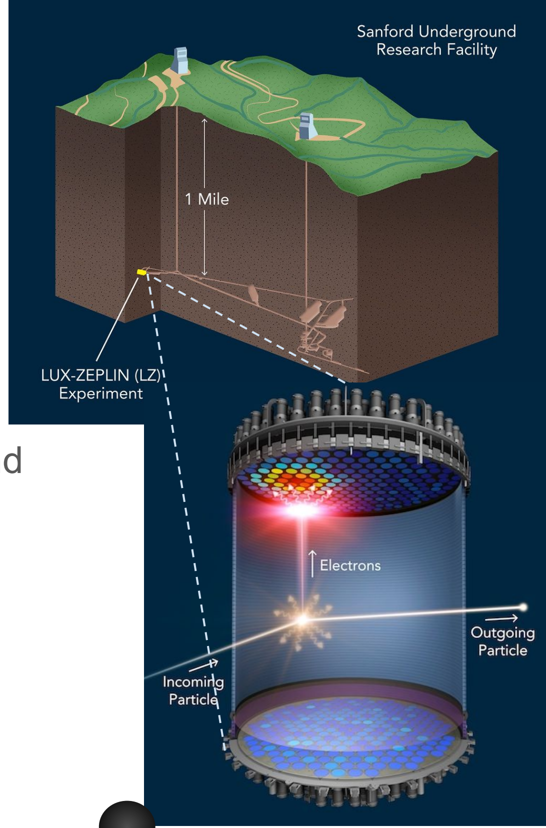
I.A Quick overview of DM search experiments

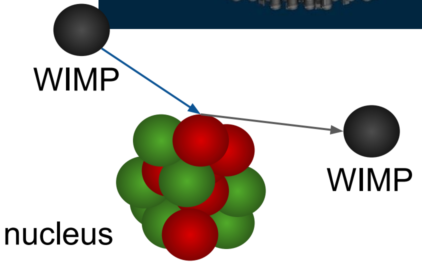
Quick Overview of the LuxZeplin Experiment

- The *LuxZeplin Experiment* (LZ) is looking for dark matter via direct measurement
- The experiment separates between dark matter interactions (**signal** events) and other interactions (**background** events) from known sources which mimic them
- By design it mitigates the background events, but the challenge is to have selection criteria which balance getting as many signal events with as few background events as possible
- Sophisticated tools are being developed to optimize the signal to background discrimination and range from high quality components to accurate reconstruction algorithms
- Simulations are a key component for these analyses (needle in a haystack problem)
- This talk is about the simulations that empower the dark matter search analyses at *LZ*

The LuxZeplin Experiment

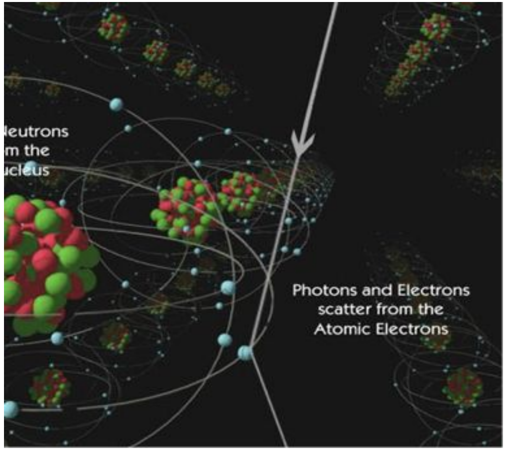
- Earth is moving through the Dark Matter halo of the Milky Way
- We are looking for an interaction between a WIMP and a heavy nucleus in the sensitive detector
- A WIMP would interact primarily with the nucleus in the liquid and produce scintillation/light in it (other models suggest strong interactions of WIMPs and electrons but are not considered here)
- Detector is deep underground (1 mile) to block particles coming from to space that could fake a WIMP interaction
- The experiment is shielded by design, to mitigate other types of radioactive decays in the mine producing high energy particles that might reach the detector from entering into our data (though some contamination is inevitable)
- In addition there are 2 layers of veto detectors near the edge regions
  (Skin - anticoincidence detector / Outer Gd scintillator)



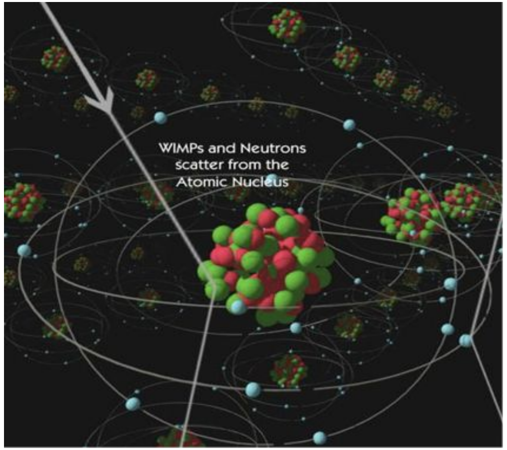


Possible Interactions and Observable Quantities

- An incoming particle can interact:
  - Electromagnetically with electrons in the outer shells, call this **Electron Recoil**
  - Non-electromagnetically off a nucleus in the liquid, call this **Nuclear Recoil**



- DM detectors would exploit **2 independent measures of energy** of the interaction (one coupled to eR, other to nR)
- Also have multiple channels for each
- The magnitude, proportions, and time-delays of these measurements allows to estimate: **energy, charge, position**



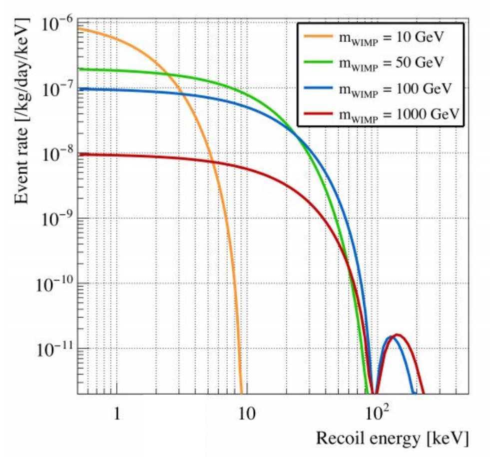
I.B The DM search challenge

Signal and Background Sources

Signal (WIMP - nR)

- Interaction via nuclear recoil
- Dominant S1 signature
- Smaller S2 signal
- Very few WIMPs are expected to interact in the detector, and those that do are expected to deposit very little energy

Number of events

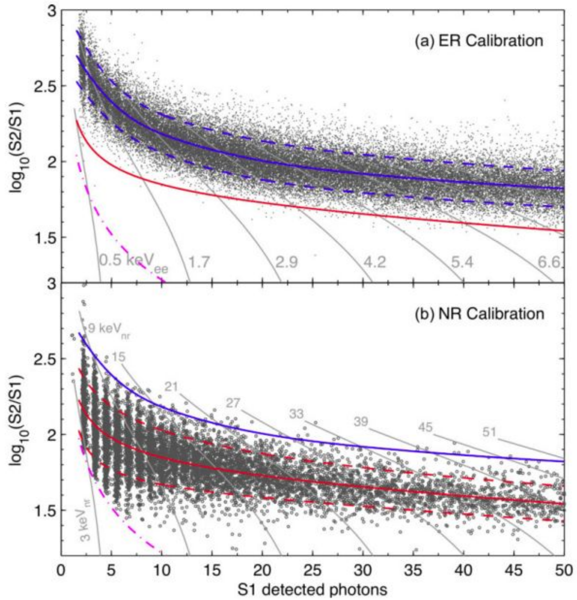


Backgrounds (eR/nR)

- From Cosmic Sources (mostly shielded):
  - Muon induced neutrons
  - Cosmogenic activation (eR)
  - Solar neutrinos (eR)
  - Atmospheric and supernova neutrinos (nR)
- Radioactive Contaminants that Decay in the Detectors (Radiogenic):
  - Mostly U238/Rn222 chain affecting detector components, and Xe
- Mismeasured Events:
  - Mixed Events: multiple particle interactions in same “event”
  - Detector/DAQ Malfunctions (?)
  - We should be able to avoid all these!

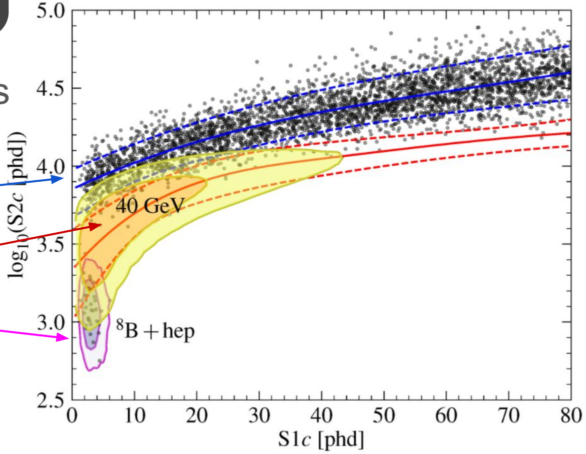
Discriminating Between Signal and Background

- Place radioactive sources near/inside the detector to create Calibration Data to understand the S1 and S2 relation:
    - Electron Recoils (e.g. CH_3T, ^{22}Na, ^{83m}Kr, ^{131m}Xe)
    - Nuclear Recoils (e.g. DD, AmLi, AmBe, ^{88}YBe)



- **Problems:**
    - The tails or *overlapping* regions are complicated
    - Interactions occurring near the edges of the detector tend to be poorly measured
    - Good quality and fiducial selection criteria based on a good understanding of the detector response are crucial

- Once the response of both eR and nR is well understood (i.e. response is calibrated and bands are fully measured) we can use them to discriminate a set of *background events*
    - Simulated background events
    - Simulated ^8B solar neutrinos and ^3He+p nR
    - 40 GeV WIMP mass contour

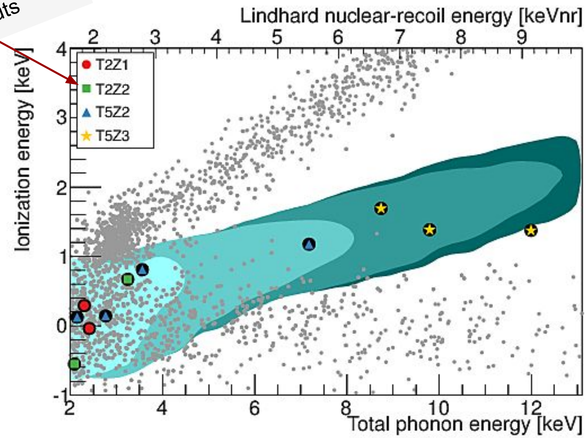


Quick Outline of a Dark Matter Search Analysis

- Goal is to observe excess WIMP events in a well understood background, or else measure the *sensitivity limit* (to a high confidence level claim that there is nothing to be observed above a cross section)
- Select good quality events
- Then select subsample of wimp-like events, by simultaneously maximizing the signal while minimizing the background number of events
- By better understanding what backgrounds and signal look like, we can help create/tune measurement techniques that allow us to better separate the two types
- This begins with making a model of each, which is highly driven by simulations

State of the Art Possible Outcomes

Events passing all cuts

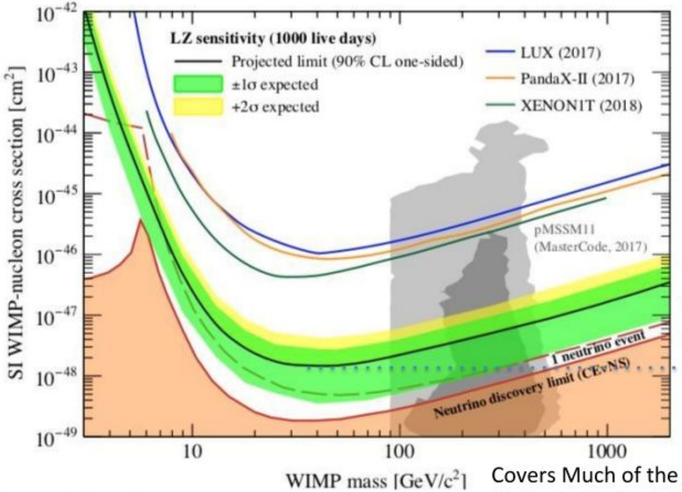


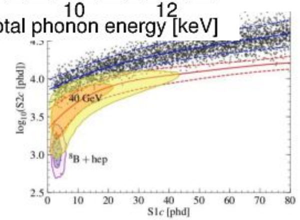
Scenario A:

Observe *excess* events giving the first *possible* direct observations of DM (becomes a *claim of discovery* with at least 5σ)

Scenario B:

Constraint WIMP-nucleon cross section limits





The best sensitivity for 1000 days, 5.6 T exposure at 40 GeV/c^2 = 1.4×10^{-48} cm^2

Covers Much of the Parameter Space from Current Limits to the Neutrino "Floor"

https://link.aps.org/doi/10.1103/PhysRevD.101.052002

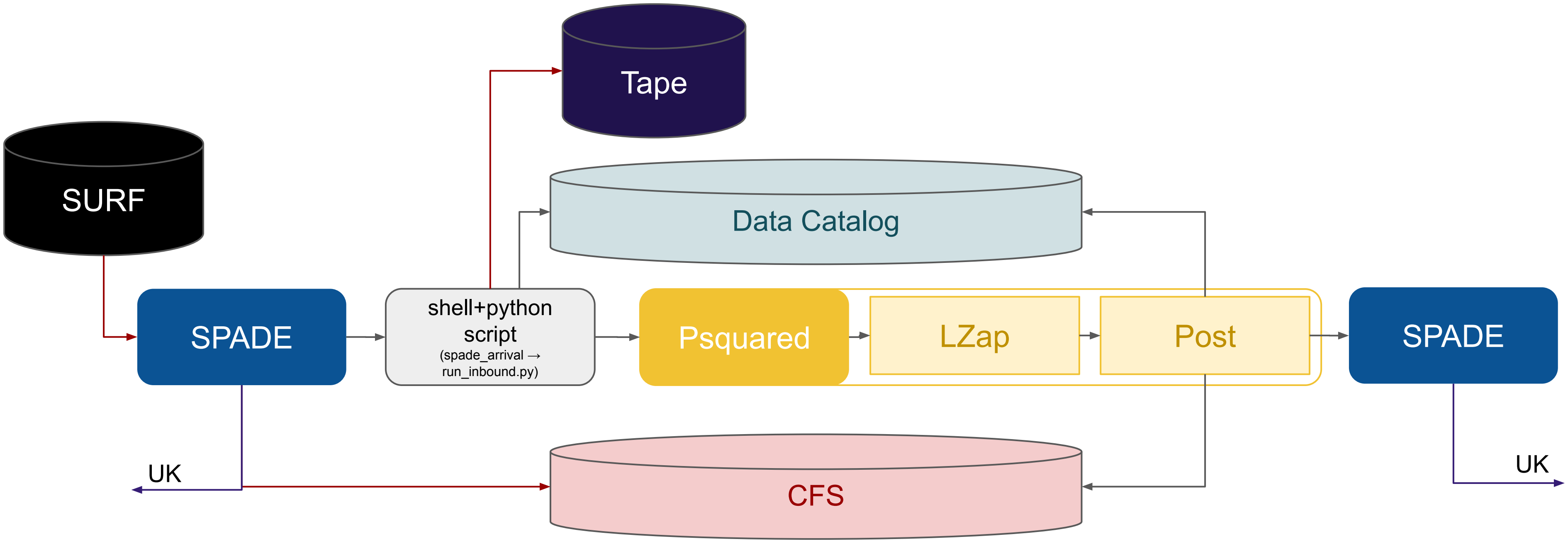
II. The Use of Simulations and ML in DM Search Experiments

Summary of The Use of Simulations and ML in DM Search

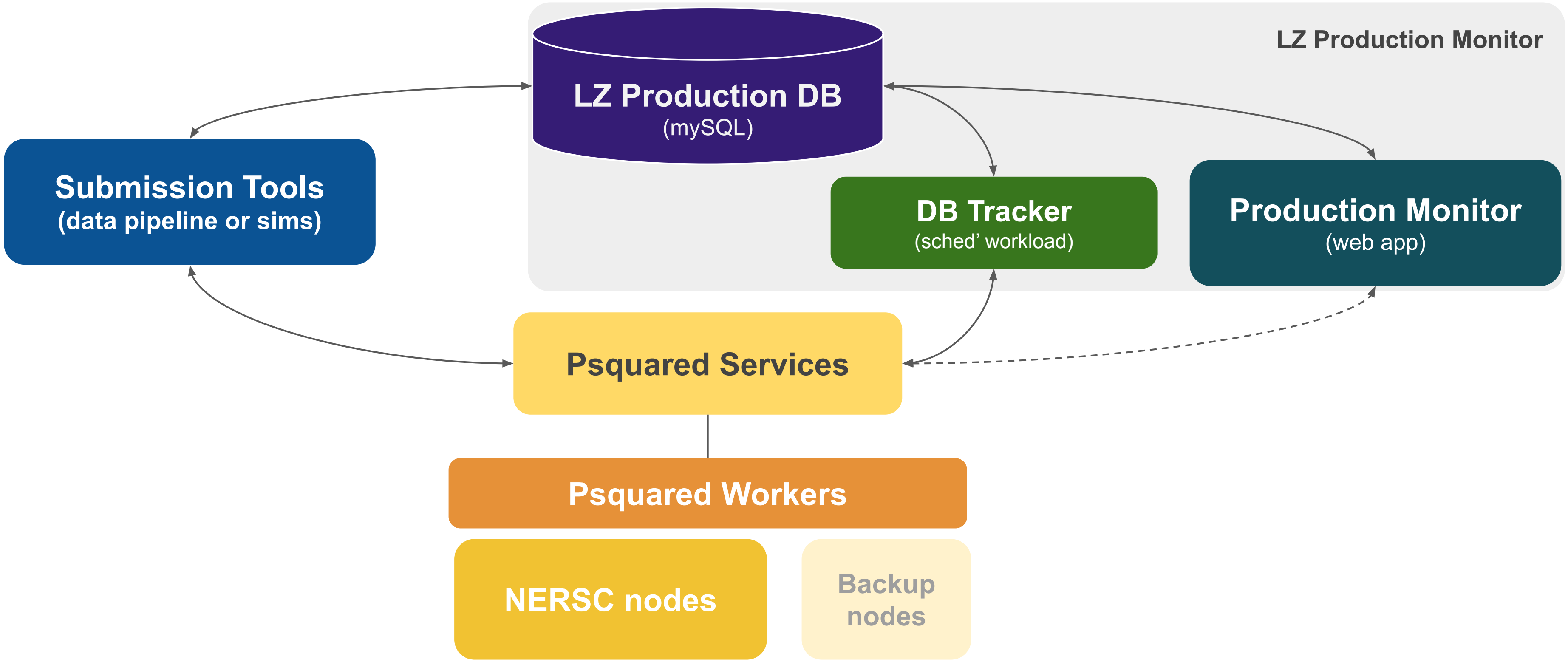
A. Automated processing pipeline
B. Simulations pipeline
C. A needle in a haystack: classification methods
D. Assembling the haystack: log-likelihood fitting of the background model
E. Best estimates of *physical* quantities from raw electronic signals: optimal filtering and more

II.A Automated processing pipeline

Prompt Processing Overview



Pipeline Backend and Monitoring Components



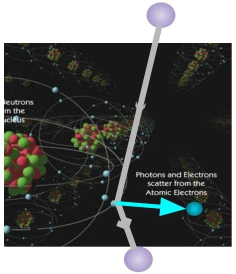
Pipeline Processing Metrics

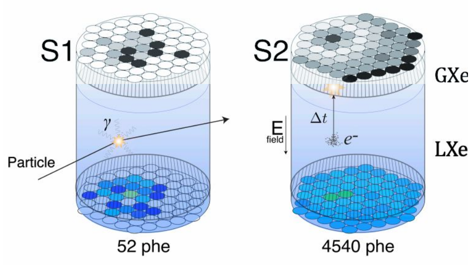
Raw data	DM	Cal	**RQ data**	Nominal RQs DM	Nominal RQs Cal
raw data size / hr [GB]	143	360	RQ data size / hr [GB]	14	36
raw data size / run [GB]	1140	2880	RQ data size / run [GB]	114	288
raw data size / day [GB]	3421	8640	RQ data size / day [GB]	342	864

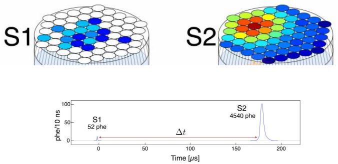
Metric	**Quantity**
Total output file size [PB]	0.295
Total CPU Hrs (2021)	4M
Number of runs/samples	4k
Number of jobs	200k
Average comp. time per event	3.5 s/ev (WS) 6-8 s/ev (calib)

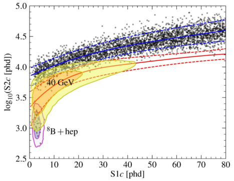
II.B Simulations pipeline

Overview of the Detector Response Simulation







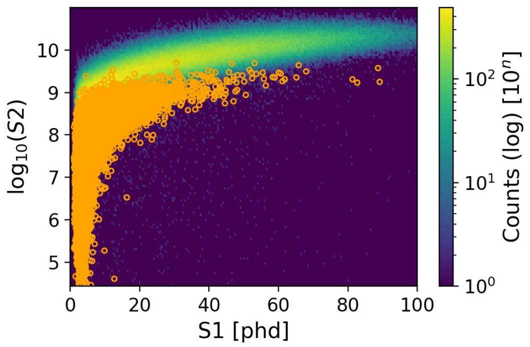


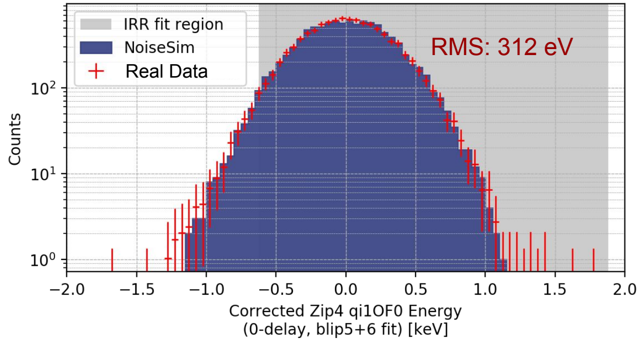
Particle Generator → Scattering and materials → Energy deposition → Light and Charge Yields → Photon Propagation → Detector Readout → Analysis event reconstruction

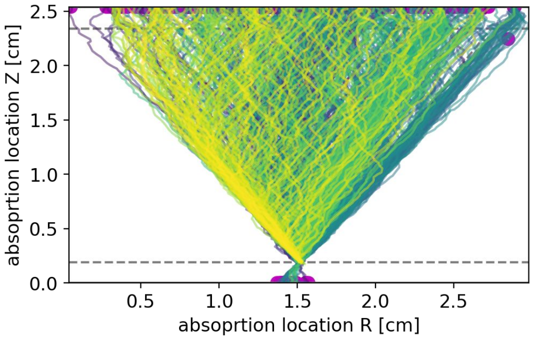
Light and Charge Yields → Charge propagation → Recombination → Photon Propagation

Simulation Productions Scale

Metric	Quantity
Total output file size [PB]	0.6
Total CPU Hrs (2021)	6M
Number of runs/samples	6k (~1.2k per prod. ver.)
Number of jobs	500k
Average comp. time per event	0.03 s/ev (BACC) 0.002 s/ev (LZLAMA)

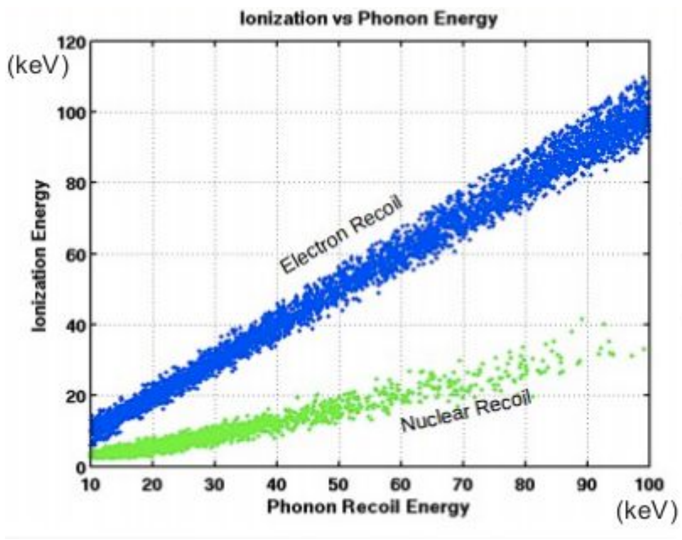


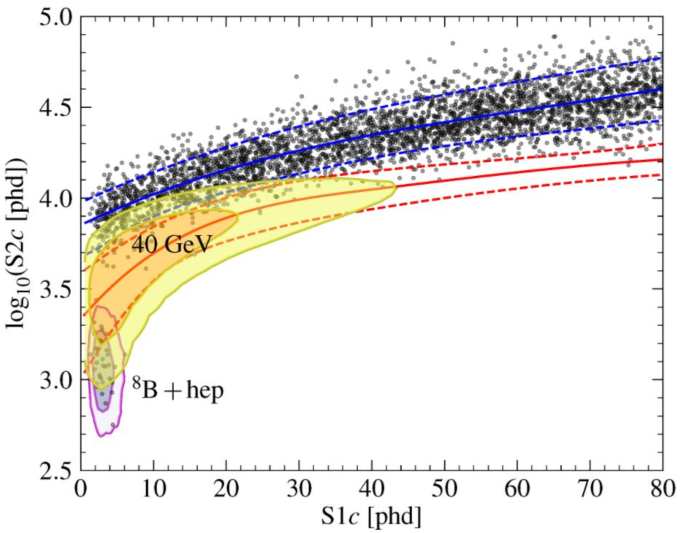




II.C Needle in a haystack: classification methods

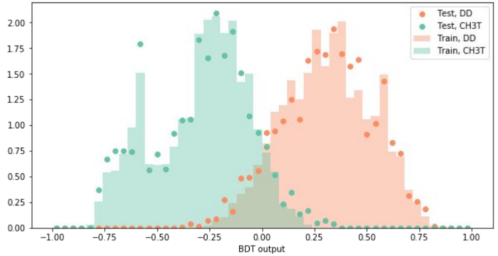
Classification Methods in DM Searches

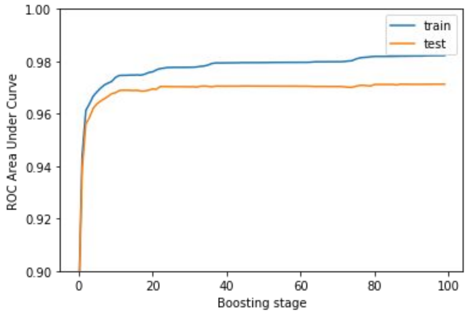


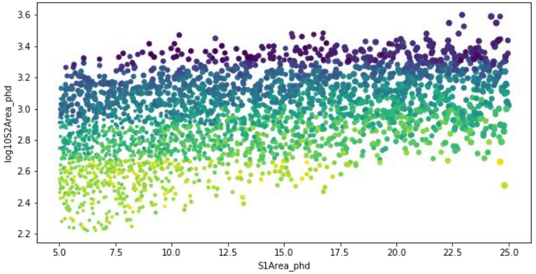


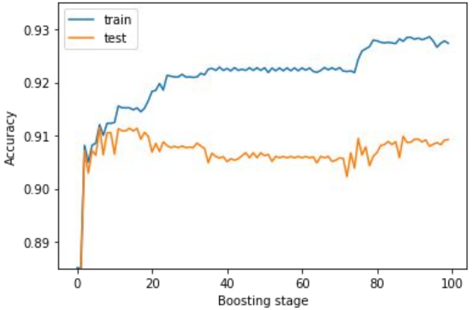
Boosted Decision Trees (BDTs)

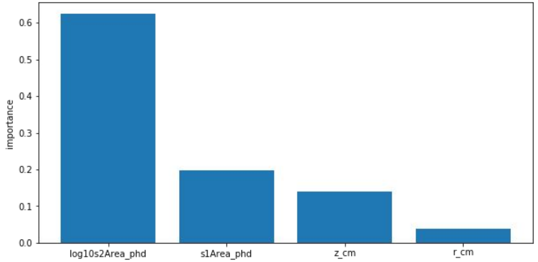
- AdaBoost-SAMME
  - Estimators (trees): 10
  - Max depth of each tree: 3
- 70/30 test/train
- Low energy focused, weighted data (for balancing datasets)
- K-fold cross validation:
  - Accuracy: 0.911 (+/- 0.018)
  - ROC area: 0.971 (+/- 0.008)

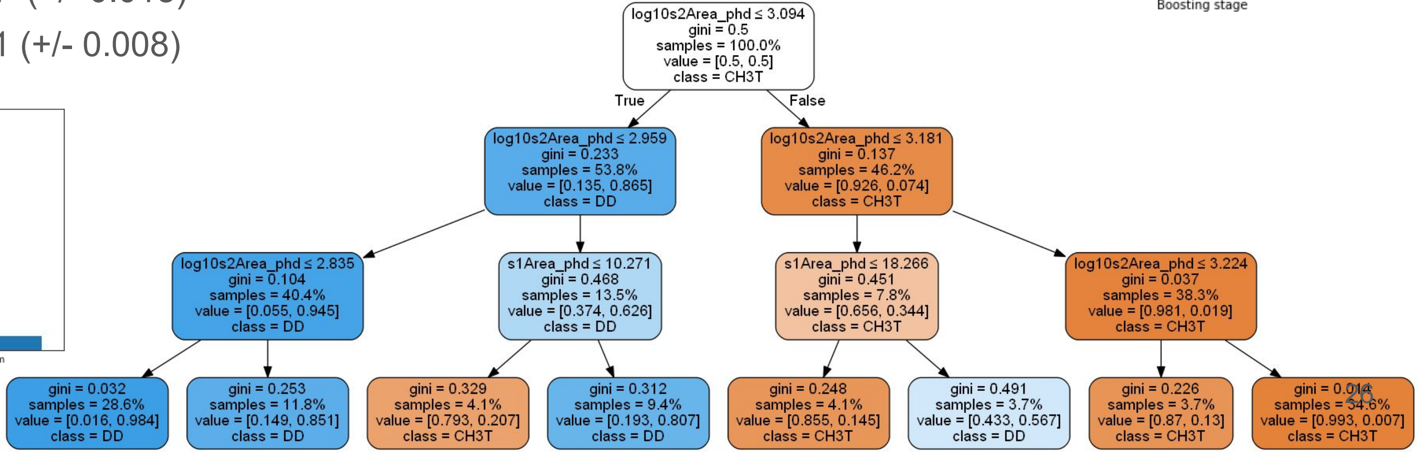






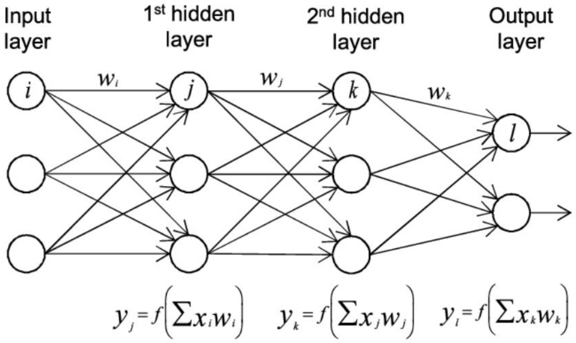


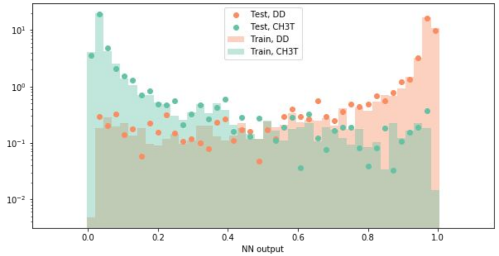


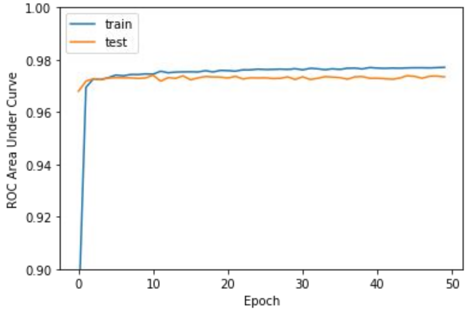


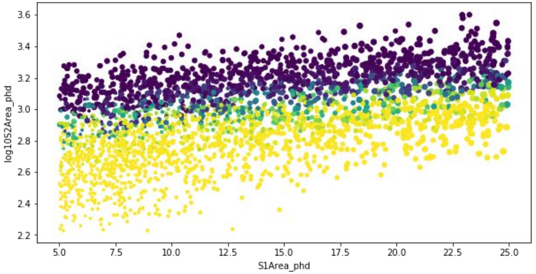
Feed-Forward Neural Network

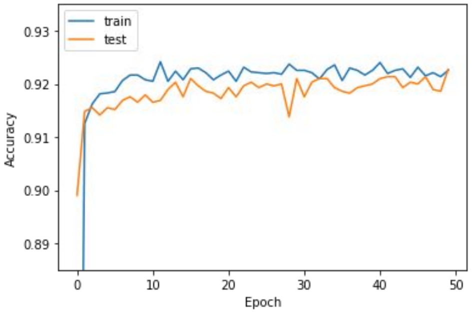
- 4 inputs → 10 node fully connected hidden layer → 3 node fully connected layer → final output layer
- Loss: mean squared error
  Optimizer: NADAM
- K-fold cross validation:
  - Accuracy: 0.914 (+/- 0.014)
  - ROC area: 0.977 (+/- 0.002)

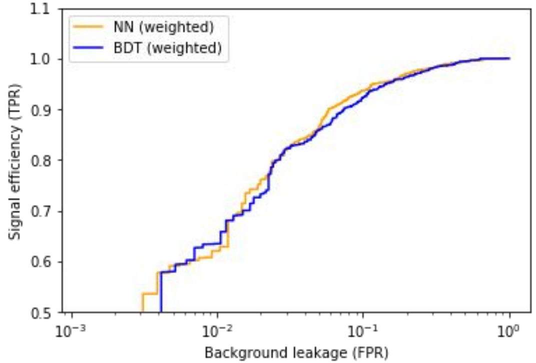


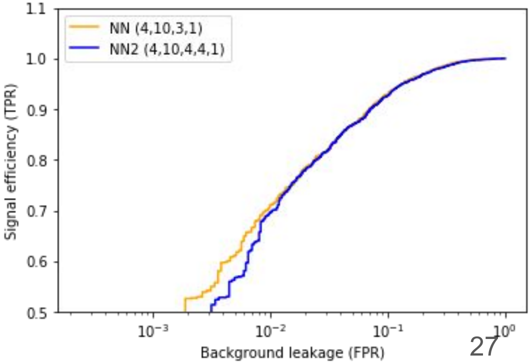








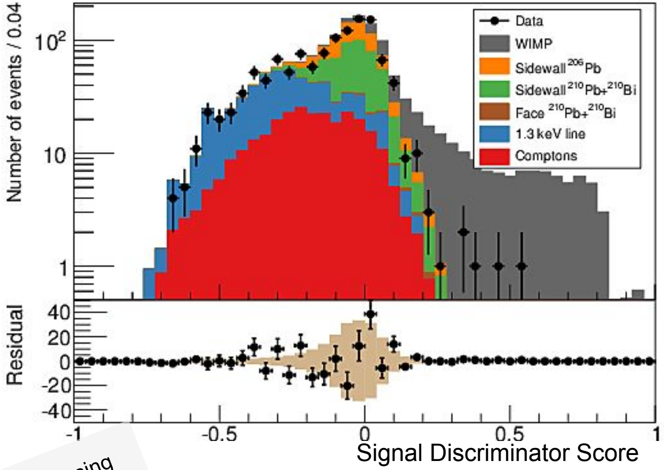




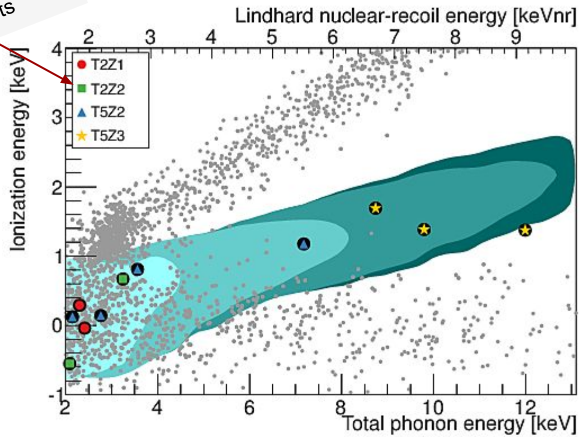
Dominant Backgrounds and Previous Analyses

SuperCDMS Collaboration. Phys. Rev. Lett. 112, 241302 (2014)

- Our current background model is based in previous analyses, where the dominant backgrounds were expected to come from 4 types of interactions:
    1. Nuclear Recoils from Lead (^{206}Pb) Contaminants
    2. Electron Recoils from Lead (^{210}Pb, ^{210}Bi) Contaminants
    3. Electron Recoils from Germanium activation (*1.3 keV line*)
    4. Cosmogenic Electrons and Photons (labelled as *Comptons*)



Events passing all cuts



- The challenge is that nuclear recoils are an irreducible background since they look like our signal, and electron recoil can look like our signal when they are mismeasured causing them to look like nuclear recoils
- The final, “optimized selection” resulted in a *handful of WIMP-like events*. They were inspected and determined to be mostly mismeasured or noise-dominated events

II.D Assembling the haystack:
log-likelihood fitting of the background model

Assembling the Background Model

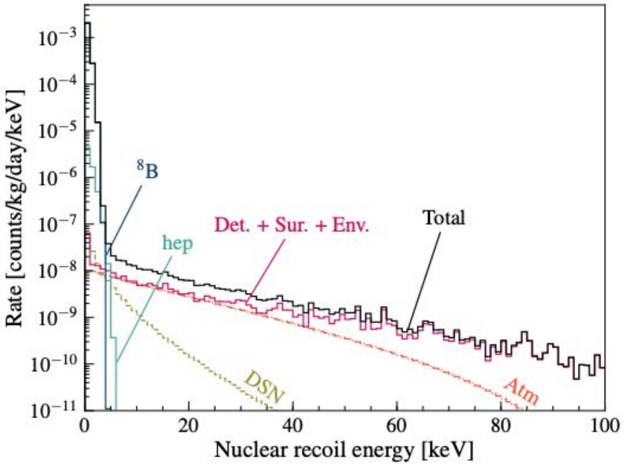
- Background Model is $f(x,\theta) = c_1 f_1(x) + c_2 f_2(x) + c_3 f_3(x) + \ldots + c_n f_n(x)$
- Create negative log-Likelihood of Background Model given data (using RooMinuit) to do fit

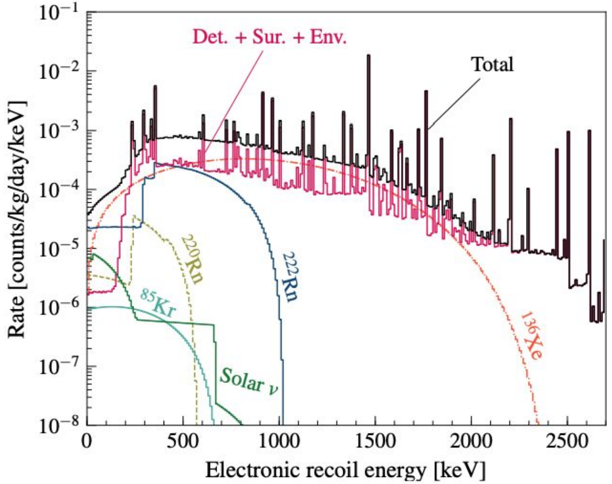
$$-\log L(\vec{x}|\theta) = -\sum_i \log f(\vec{x}_i|\theta)$$

- Don't add explicitly the ~1.1k samples (pdfs), instead we add them up with some predefined normalizations into major contributions
- Then we do the background model fit

LZ Background Model

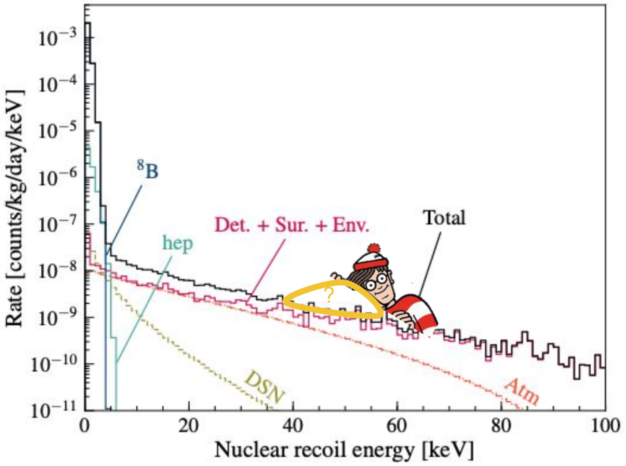
- Our current background model is dominated by radiogenic contaminants in general (except for a few regions/bins)
- If we can precisely match the contributions of each component to our observations then we can trust that excess events may be associated with dark matter
- We thus simulate events for each background source (and its location) independently and then we have tools to stitch/fit them together based on our observations
- This model (in combination with the separation technique) is the central piece of our dark matter searches

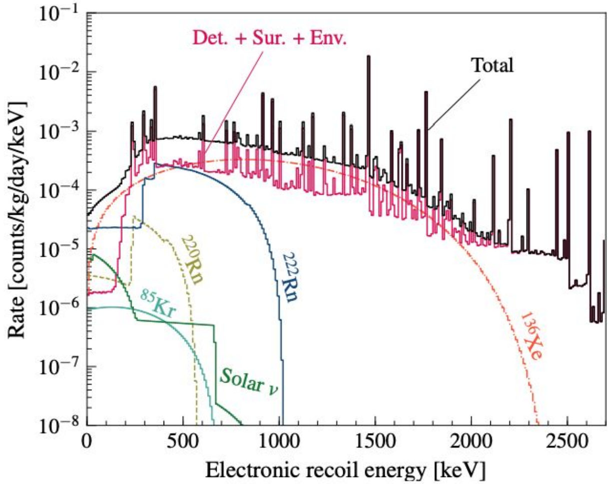




LZ Background Model

- Our current background model is dominated by radiogenic contaminants in general (except for a few regions/bins)
- If we can precisely match the contributions of each component to our observations then we can trust that excess events may be associated with dark matter
- We thus simulate events for each background source (and its location) independently and then we have tools to stitch/fit them together based on our observations
- This model (in combination with the separation technique) is the central piece of our dark matter searches

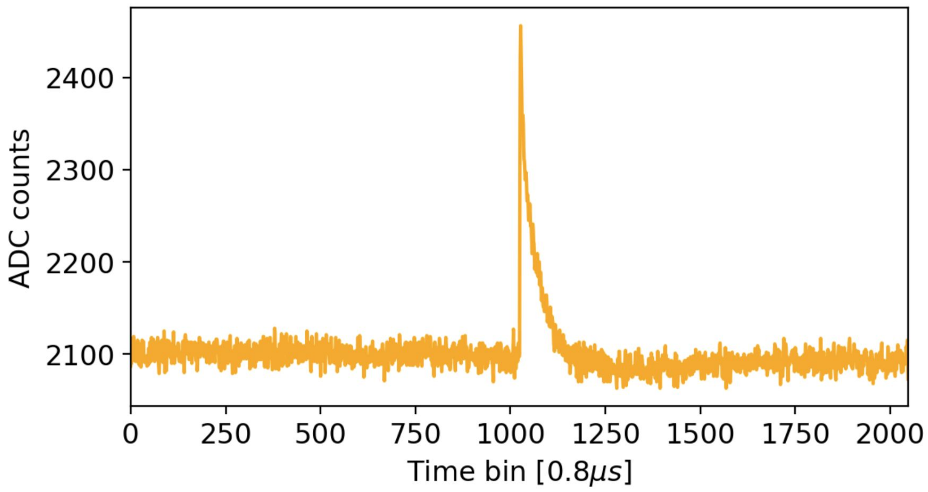




II.E Best estimates of *physical* quantities from raw electronic signals: optimal filtering and more

Simplistic Scenario: Signal over Noisy Background

Estimate the amplitude → energy



Max?
Integral?
Time-domain fit?
Frequency-domain fit?

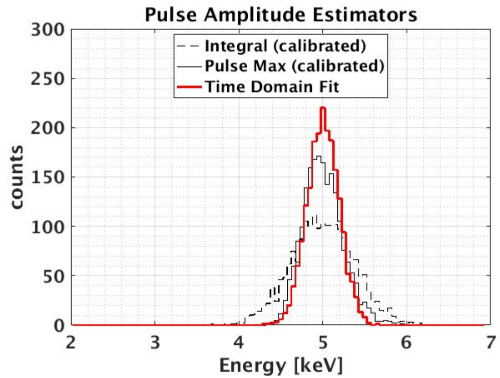
Simplistic Scenario: Signal over Gaussian Noise

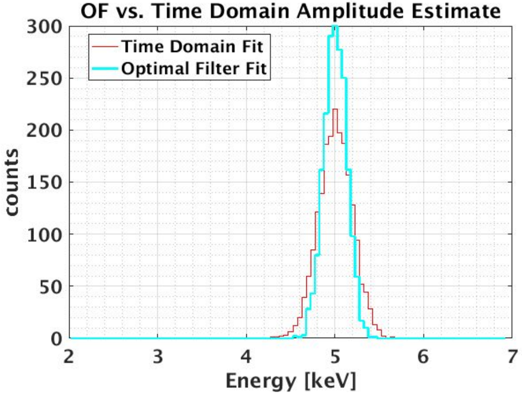
- *Integral*: Vulnerable to noise, no goodness of fit parameter (χ^2)
- *Pulse Max*: Vulnerable to noise, uses only 1 data point, no χ^2
- *Time domain fit* (expected signal shape, and fixed t_0):
  - Fit functional form: $$\underbrace{S(t)}_{\text{signal}} = \underbrace{a}_{\text{pulse amplitude}}\underbrace{A(t)}_{\text{known template}} + \underbrace{n(t)}_{\text{gaussian noise}}$$

    Calculate a time-domain χ^2 : $$\chi^2_{TD}(a) = \sum_{k=0}^{N} \frac{(S_k - aA_k)^2}{< n_k^2 >}$$

    minimize for best fit: $$a\hat{T}D = \frac{\sum_k \frac{S_k A_k}{<n_k^2>}}{\sum_k \frac{A_k^2}{<n_k^2>}}$$
  - **Problem**: the above χ^2 is not a proper maximum likelihood estimator if the time domain bins are correlated. Time bins are correlated unless noise is white (flat power spectrum)
    - ⇛ Since noise in our experiment is not *white* this isn't the best solution

2000 simulated iZip phonon signal+noise 5keV energy pulses

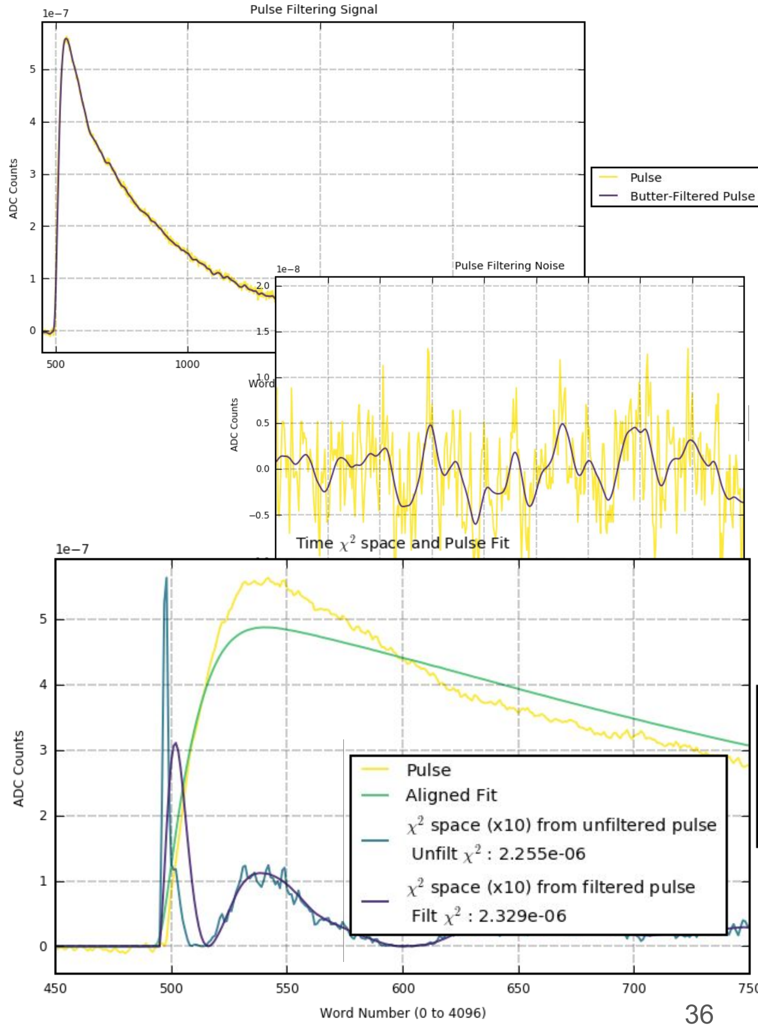




➢ $<n_k^2>$ is variance or "error bar" of kth time bin. Expect constant $<n_k^2>$ for Gaussian noise.

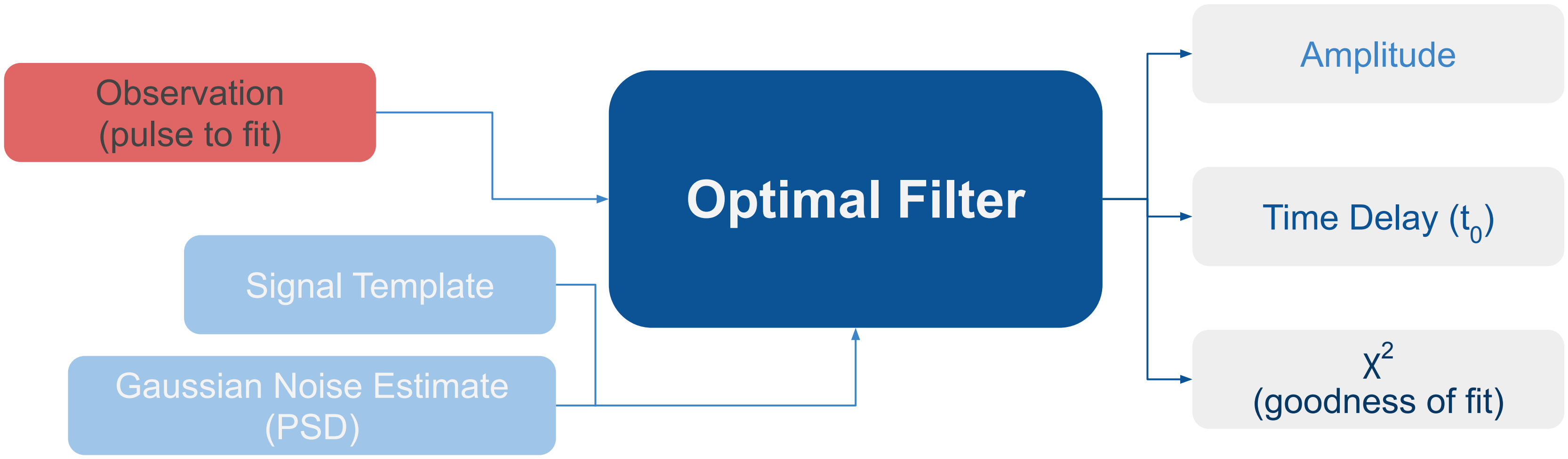
Turn to Freq Space: Filtering a Signal

- Come up with a technique for filtering-away unwanted (*noise*) components from a *pulse* and extracting the best fit from a *signal* model
- One could think about:
  - A bandpass (butterworth) filter, and take certain frequencies away from the signal
  - Cycling again, with different frequencies until you are left with a *good enough* pulse
  - Then apply the fit from the *signal model*
- Problem is that you don't want to kill-away features of your *signal* but you want to kill as much noise as possible
- What is the method for ***optimally*** crafting a filter and obtaining the best amplitude estimate?



Solution: The Optimal Filter

- Is a method to *amplify* your signal contributions through a *filter* specifically designed to de-weight the known (gaussian random) noise spectrum from your observations
- Assumes the signal has a *known shape*
- Requires an estimate of the *expected* noise, and for it to be random gaussian



Requirements for the Optimal Filter

$$v(t) = As(t) + n(t)$$

observation

fit amplitude

signal template

gaussian noise

- Let the *observation* $v(t)$ be the sum of a *signal template* $s(t)$ with gaussian random noise $n(t)$
- Go to frequency domain, by *Discrete-Fourier-Transforming* the observation, template, and noise estimate (time-domain bins → freq. n-index)
- Construct *Power Spectral Density* (PSD) of expected noise, labelled J_n which is the auto-correlated variance of each frequency bin
  - $J_n = < \tilde{n}_n^2 >$
  - Freq. domain data points are uncorrelated (i.e. noise covariance matrix is diagonal)
  - This is the feature that breaks the time-domain-fit problem!
- Construct the frequency domain χ^2, and minimize to obtain minimum χ^2 and maximum $\hat{A}$, at delay t_0

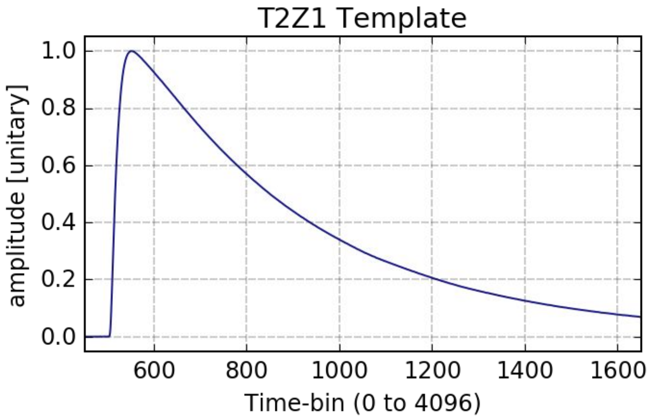
$$\chi^2 = \int_{-\infty}^{\infty} df \frac{|\tilde{v}(f) - Ae^{-j\omega t_0}\tilde{s}(f)|^2}{J(f)}$$

$$\chi^2 = \sum_{n=-\frac{N}{2}}^{\frac{N}{2}-1} \frac{|v_n - e^{j\omega_n \widehat{t_0}} \widehat{A}\, \tilde{s}_n|^2}{J(f_n)}$$

$$\hat{A} = \frac{\sum_{n=-\frac{N}{2}}^{\frac{N}{2}-1} e^{j\omega_n \widehat{t_0}} \frac{\tilde{s}_n^* v_n}{J(f_n)}}{\sum_{n=-\frac{N}{2}}^{\frac{N}{2}-1} \frac{|\tilde{s}_n|^2}{J(f_n)}}$$

Fourier transform (denoted by '~')

The Signal Template

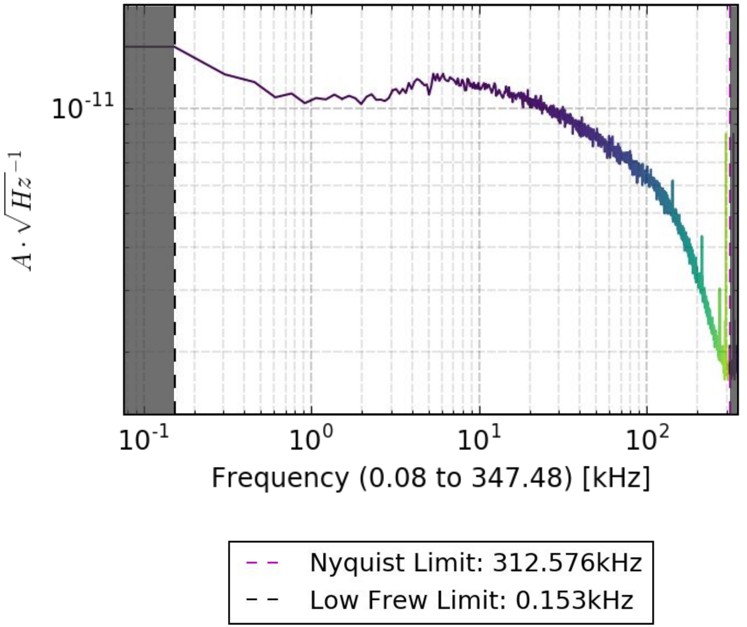


- The *signal template* is the average good quality pulse
- Typically obtained by some recipe:
  - Make a *good pulse quality* event selection (analysis dependent)
  - Select an energy that's high enough above noise, but without pulse saturation (i.e. charge: 80-120keV)
  - Extract and align the selected pulses (different alignment methods can be followed)
  - Normalize the pulses (with the pulse integral)
  - Get average pulse (average value at every bin)
  - Keep the 80% best time-domain χ^2 against the average pulse
  - Normalize again, but this time with the Optimal Filter calculation, using the previous average as the template
  - Calculate the average of the OF normalized pulses, and divide by its maximum value
  - Further steps followed for residual template and non-stationary covariance matrix

The Expected Noise (PSD)

- Describe the distribution of power across the frequency components within the discrete FFT decomposition
    - Get noise sample (typically 500 *random triggers* taken before each series)
    - FFT the noise sample (and typically divide by number of bins and digitization rate)
    - Compute the variance of each frequency bin σ_n^2 (times the digitization rate)
    - The 0th and last (Nyqist freq) PSD bins are σ_n^2 and the rest are $2\ \sigma_n^2$ (because bins beyond the Nyqist frequency are the negative/symmetric image of the rest)
    - Resulting units are A^2/Hz, typically sqrt of abs is plotted

- Notice how, the time-wise length of the pulse, and the digitization time determine the frequency limits
  $dv = 1/(nbins\ dt)$



➢ The Nyqist frequency is the n/2 index

What is the Filter? How does the OF work?

Inverse FFT, gives amplitudes at any delay *t'*, look for max and get t_0

$\tilde{s}_n$ → template FFT

$J(f_n)$ → noise PSD

v_n → observ. FFT

$$\hat{A} = \frac{\sum_{n=-\frac{N}{2}}^{\frac{N}{2}-1} e^{j\omega_n \hat{t_0}} \frac{\tilde{s}_n^* v_n}{J(f_n)}}{\sum_{n=-\frac{N}{2}}^{\frac{N}{2}-1} \frac{|\tilde{s}_n|^2}{J(f_n)}}$$

Normalization Constant:

$$\Sigma_i \frac{|templ\_fft_i|^2}{PSD^2_i}$$

(AKA: signal to noise)

Filter

$$\Phi_i = \frac{templ\_fft_i^*}{PSD^2_i}$$

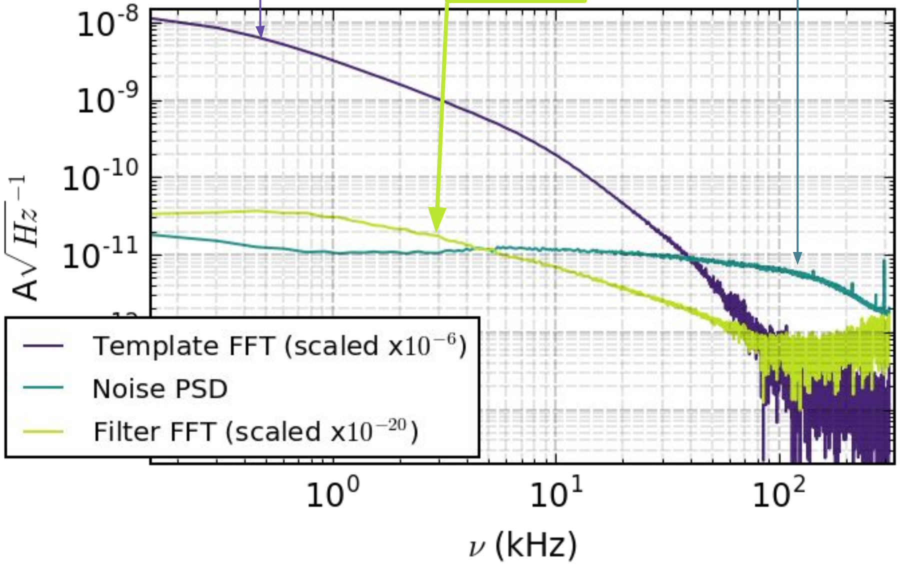
Φ_i multiplies v_i, so for a product ($\Phi_i v_i$) the resulting *amplitude* is proportional to the filter amplitude at such particular frequency → noise dominated frequencies are de-weighted

See the Filter in Action (Example 1)

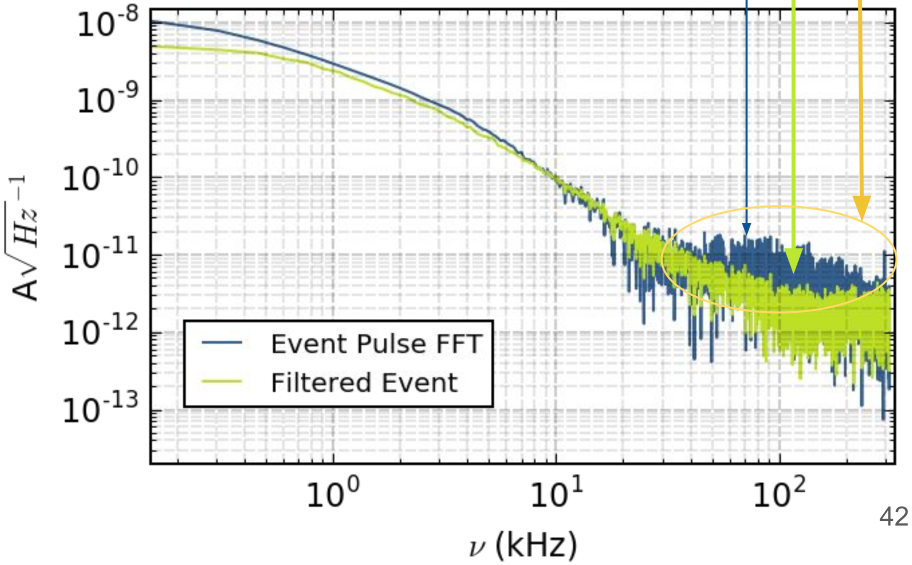
- The filter (green) is the ratio of the Template FFT (Im) and the Noise PSD

- Notice how at frequencies where noise dominates the filter has lower values, thus those frequencies are damped in the filtered event

$$\frac{\tilde{s}_n^* v_n}{J(f_n)}$$



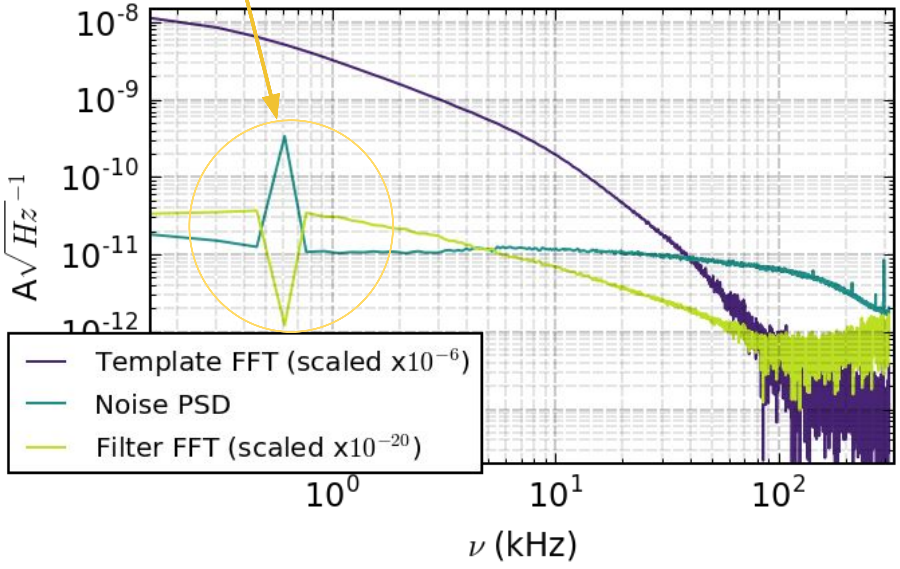
$$\frac{\tilde{s}_n^* v_n}{J(f_n)}$$



See the Filter in Action (Example 2 Artificial Noise Peak)

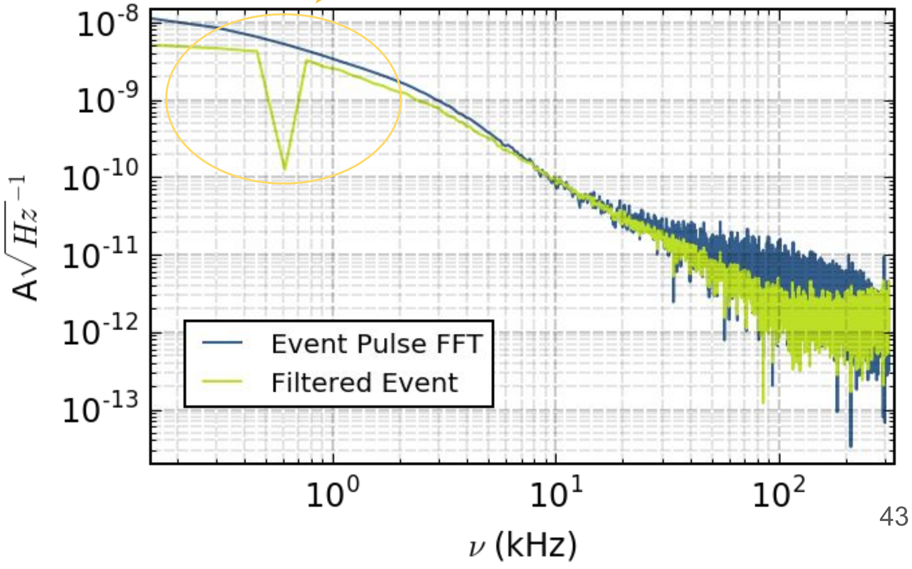
- A noise peak has been added *artificially* to the NoisePSD (i.e. a sine-wave in time-domain)
- The filter responds compensating for such frequency

$$\frac{\tilde{s}_n^* v_n}{J(f_n)}$$



- Now, the filtered event also gets that frequency damped
- This way, any noise features get 'killed' over the signal

$$\frac{\tilde{s}_n^* v_n}{J(f_n)}$$

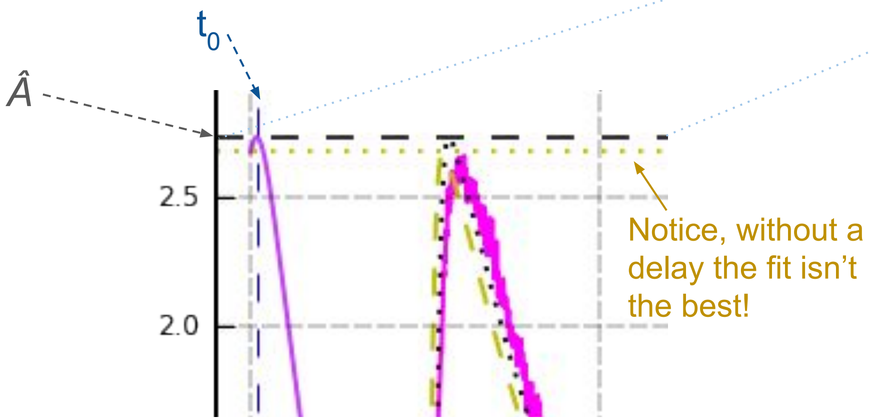


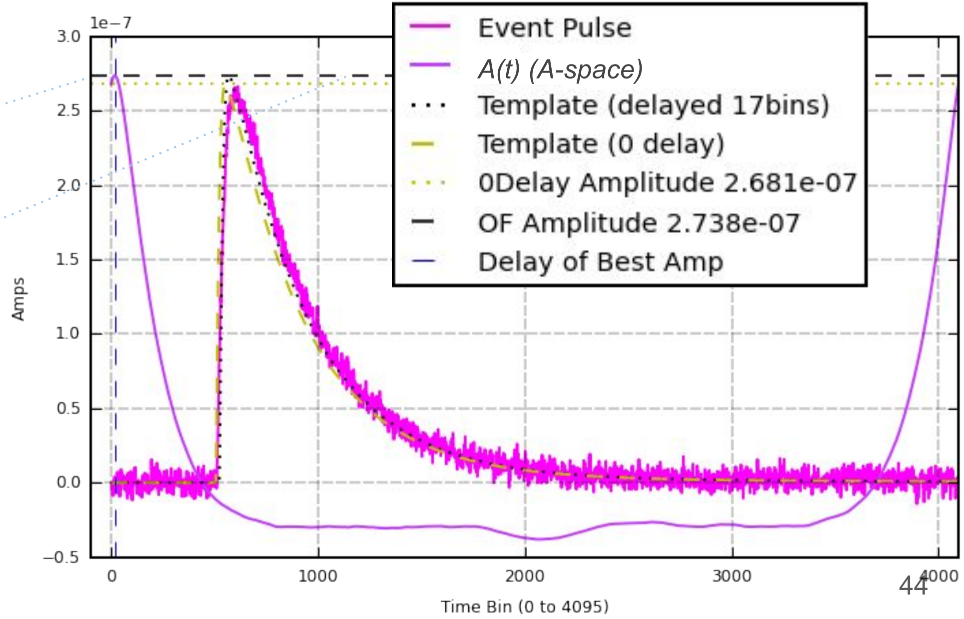
Optimal Filter Fit Example

- *A*, before being $\hat{A}$ (fixed at t_0) is just the normalized inverse FFT of the filtered event
- Upon inverse transforming you have an '*A space*' where:
  - The maximum value is $\hat{A}$,
  - Such bin-index is the delay t_0

Inverse FFT of filtered event

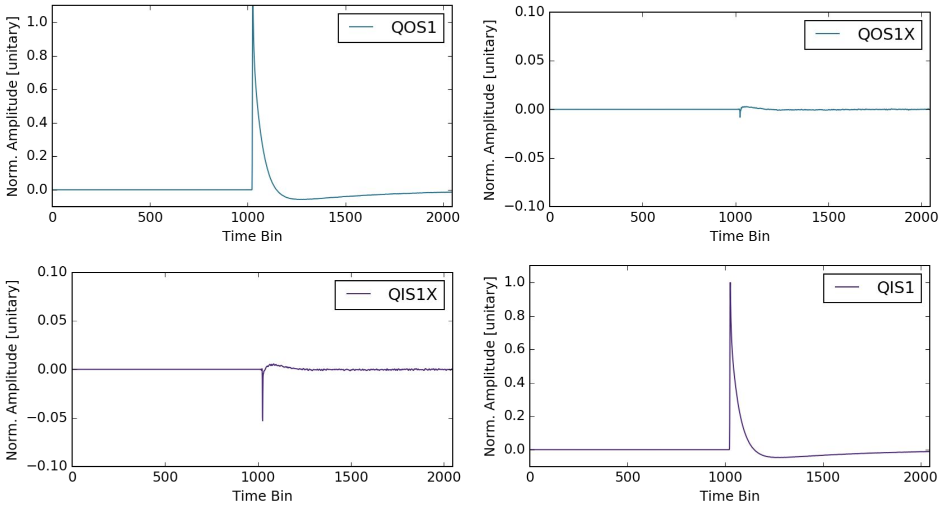
$$\hat{A} = \frac{\boxed{\sum_{n=-\frac{N}{2}}^{\frac{N}{2}-1} e^{j\omega_n \widehat{t_0}} \frac{\tilde{s}_n^* v_n}{J(f_n)}}}{\sum_{n=-\frac{N}{2}}^{\frac{N}{2}-1} \frac{|\tilde{s}_n|^2}{J(f_n)}}$$





Simultaneously Solved Optimal Filter in Charge Channels

- Charge channels (from same side) have capacitively induced crosstalk
  - ~5% in CDMS II detectors
  - ~1% on iZIPs
- Create a fit system minimizing/solving simultaneously for the amplitude and χ^2 of 2 templates in addition with the 2 crosstalk templates



$$\begin{pmatrix} S_I \\ S_O \end{pmatrix} = \begin{pmatrix} A_I & A_{Ix} \\ A_{Ox} & A_O \end{pmatrix} \begin{pmatrix} a_I \\ a_O \end{pmatrix} + \begin{pmatrix} n_I \\ n_O \end{pmatrix}$$

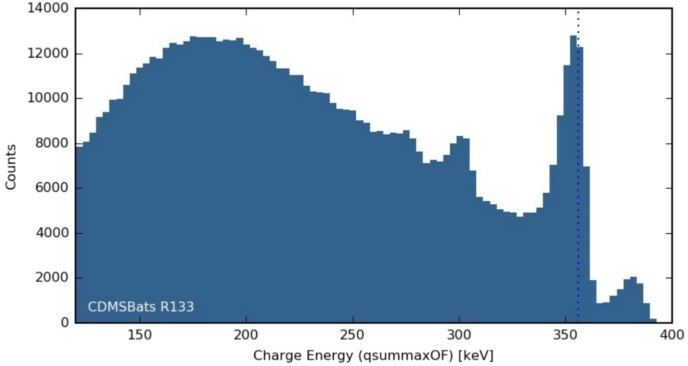
$$\chi^2(a_I, a_O) = \sum_n \left(\frac{|\tilde{S}_I - (a_I \tilde{A}_I + a_O \tilde{A}_{Ix})|^2}{J_I} + \frac{|\tilde{S}_O - (a_I \tilde{A}_{Ox} + a_O \tilde{A}_O)|^2}{J_O} \right)$$

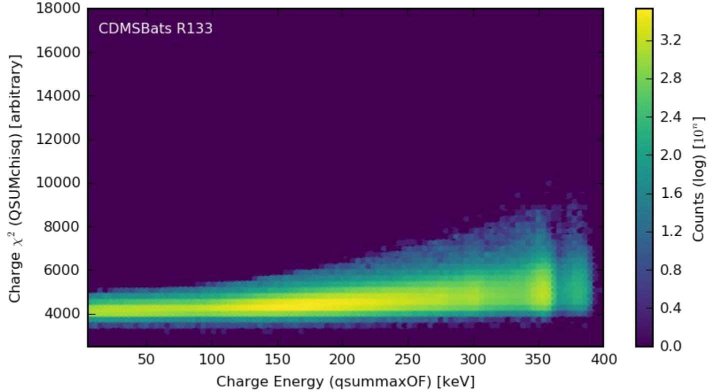
$$\begin{pmatrix} \hat{a}_I \\ \hat{a}_O \end{pmatrix} = M^{-1} \begin{pmatrix} \sum_n \mathrm{Re}\left(\frac{\tilde{A}_I^*}{J_I} \tilde{S}_I + \frac{\tilde{A}_{Ox}^*}{J_O} \tilde{S}_O \right) \\ \sum_n \mathrm{Re}\left(\frac{\tilde{A}_{Ix}^*}{J_I} \tilde{S}_I + \frac{\tilde{A}_O^*}{J_O} \tilde{S}_O \right) \end{pmatrix}$$

$$M \equiv \begin{pmatrix} \sum_n \frac{|\tilde{A}_I|^2}{J_I} + \frac{|\tilde{A}_{Ox}|^2}{J_O} & \sum_n \mathrm{Re}\left(\frac{\tilde{A}_I^* \tilde{A}_{Ix}}{J_I} + \frac{\tilde{A}_O^* \tilde{A}_{Ox}}{J_O} \right) \\ \sum_n \mathrm{Re}\left(\frac{\tilde{A}_I^* \tilde{A}_{Ix}}{J_I} + \frac{\tilde{A}_O^* \tilde{A}_{Ox}}{J_O} \right) & \sum_n \frac{|\tilde{A}_O|^2}{J_O} + \frac{|\tilde{A}_{Ix}|^2}{J_I} \end{pmatrix}$$

Charge OF Amplitude and χ^2 in SuperCDMS Data

- Look at example R133 Barium energy and χ^2 (with good some quality cuts applied*)
- Charge measurements behave in more stable manner (w.r.t. energy dependence)
  - Notice sharper γ lines
  - Notice χ^2 not going too far from the number of d.o.f.
- The presence, absence, or ratio of the charge channels can also be used to generate rough fiducial cuts





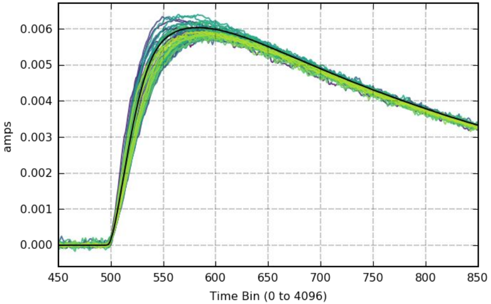
Methods Beyond Optimal Filtering

- Can we extend to more than one signal template?
  - Yes, 2nd template (residual) is related with position and saturation!
  - Further? → Principal components analysis, connect the PCs with the signal templates, and can use the PCs to project signal estimates!

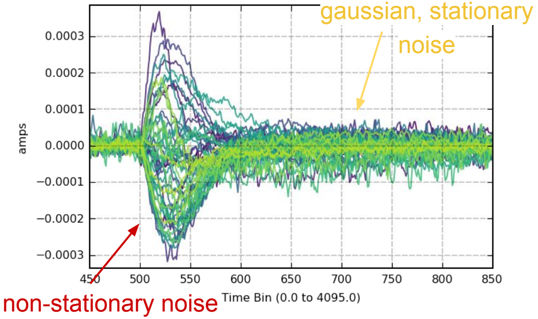
- What if noise is cross-frequency correlated? → Non-stationary Optimal Filter

- Can we implement a CNN frequency space which would convolve based on the signal-correlated-frequencies?

Non-Stationary Optimal Filter

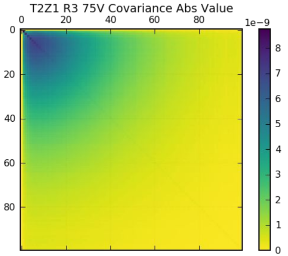


- Recall how the phonon OF χ^2 depends (semi) quadratically on energy , that is mostly due to saturation effects in the TES
- Also recall how pulse shape varies a lot with the recoil location
- In the eyes of the Optimal Filter, this is *interpreted* as non-stationary noise
- Can modify the Optimal Filter and treat these pulse/detector features as a different type of noise accounted for in the filter



- Redefine the expected variation as the cross-correlation of all frequencies; where the diagonal is the equivalent autocorrelation function

$$J(\nu) = \langle |\tilde{n}(\nu)|^2 \rangle \mapsto \tilde{V}(\nu, \nu') = \langle \tilde{n}^*(\nu' - \nu)\tilde{n}(\nu) \rangle$$



  - Thus, we have an additional input, the Covariance Matrix, whose diagonal is the PSD

- The χ^2 is now: $\chi^2(a, t_0) = \sum_{ff'} [\tilde{S}^*(f)e^{2\pi i f t_0} - a\tilde{A}^*(f)][\tilde{V}(f,f')^{-1}][\tilde{S}(f')e^{-2\pi i f' t_0} - a\tilde{A}(f')]$

III. Conclusions

Conclusions

- Dark matter searches are empowered by ML and other sophisticated computational techniques to squeeze for optimal results
- Experimental data is processed live through streamed pipelines
- Large volumes of simulated data are needed for model training and to challenge detector understanding
- Classification models are a square-1 start for the background separation
- Sophisticated background models are assembled with thousands of simulated samples, the best fit proportion tells about the contributions to the spectrum
- Best estimate extraction algorithms are essential for good quality observations, the challenge tends to be on the low-amplitude/high noise regime
- We have often started from the *physics* to find the best *tool* but we are now challenged by studying the *optimal tools* which empower the discovery of new physics

Acknowledgements

For funding all or parts of this work

SLAC NATIONAL ACCELERATOR LABORATORY

U.S. DEPARTMENT OF ENERGY
Office of Science

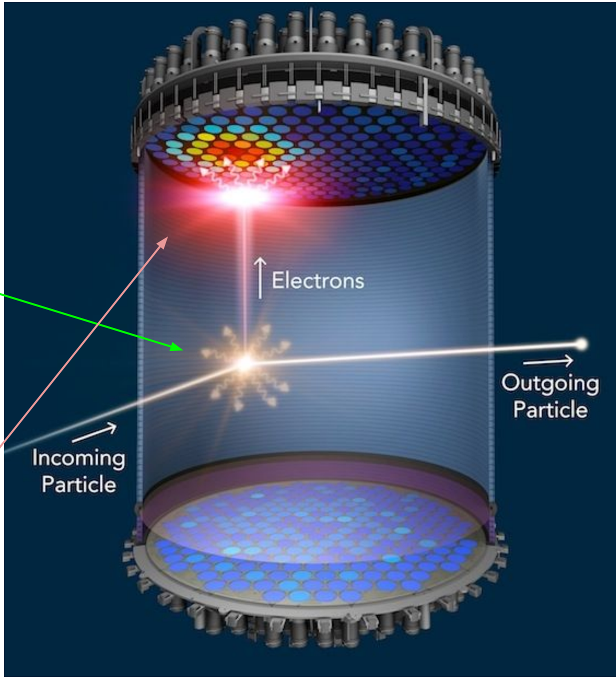
LZ

SUPER CDMS

Did not fund this work,
but is current employer

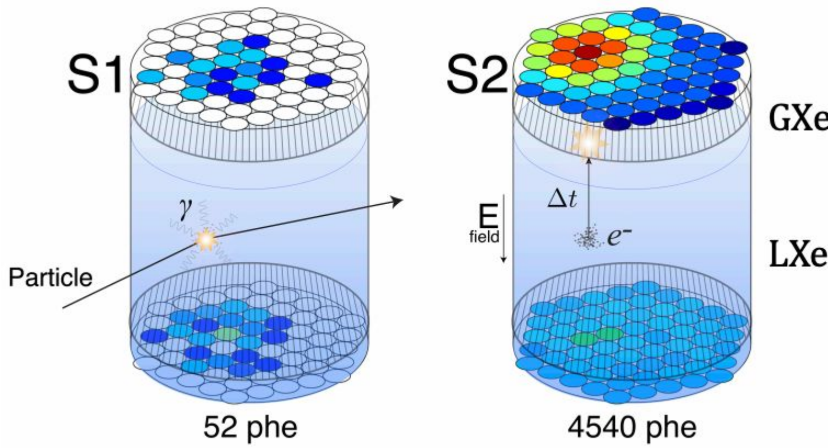
Interaction Signals in the Detector: S1 and S2

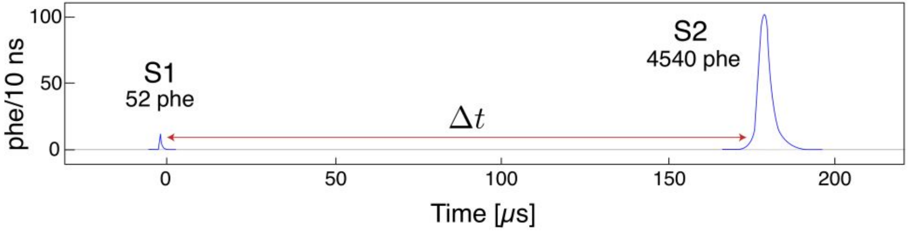
- After a particle interacts we record 2 types of responses:
  - Prompt Scintillation S1:
    - Deposited directly by the recoiling particle into the gas
    - A Xe nucleus is excited, bound to another, and emits light (scintillation) in the process
    - All happens quickly 3/27ns (singlet/triplet)
    - We call this signal S1
    - Neutral particles produce more S1
  - Secondary scintillation S2:
    - Electrons are liberated from the atom(s)
    - Eventually they recombine, excite, and cause scintillation again in a similar process
    - But we pull them to the top with an electric field
    - After drifting in the chamber, recombine in the Xe gas at the top
    - Call this signal S2
    - Charged particles produce more S2
- Every particle interaction results in both types of energy deposition, the proportion varies with the interaction type (particle type, charge, and energy)



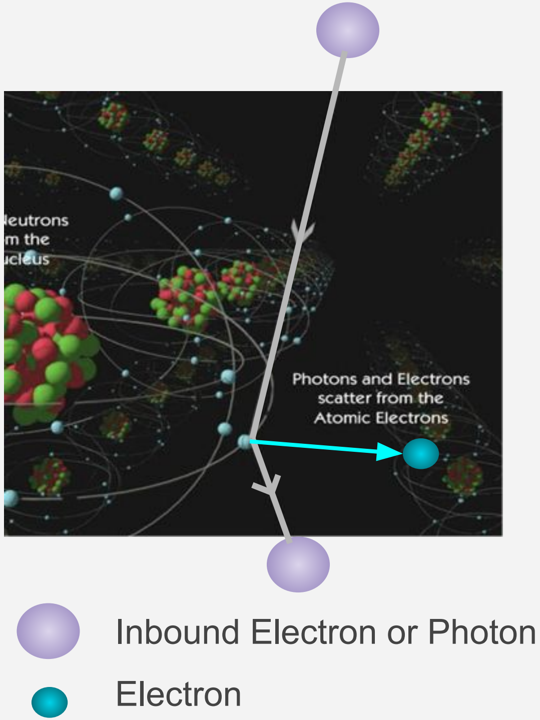
Using S1 and S2 to Observe Interactions

- Photons travel through the chamber bouncing on the highly reflective inner shell (Polytetrafluoroethylene - PTFE) until reaching the top/bottom
- Photomultiplier tubes (PMTs) collect them and produce an electric readout
- The integrated area of each pulse is proportional to the photoelectrons (phe) and in turn proportional to the energy deposition (of each S1 and S2)
- The arrays provide radial position by concentric pattern
- The delay between S1 and S2 gives the depth position
- Triggering on, and finding the signals to then measure the parameters is also a challenge (i.e. pileups, faint signals, noise, etc)



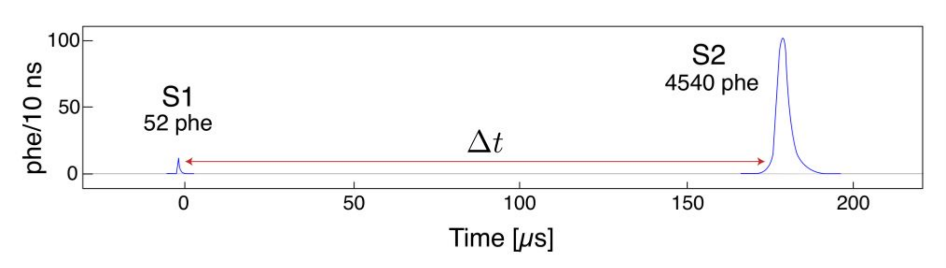


Details About Electron Recoils



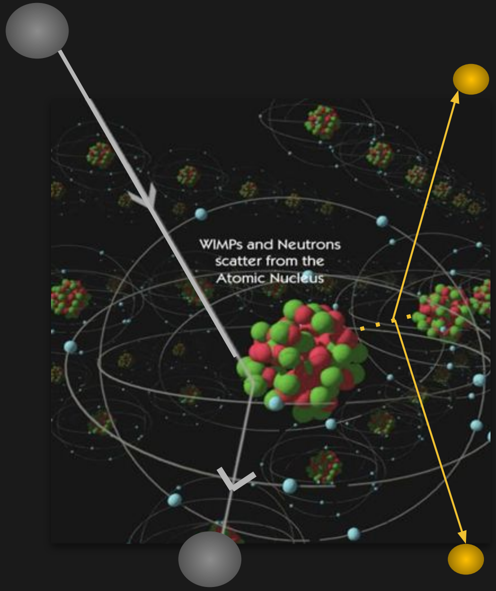
- Recall that the proportion of S1 (prompt scintillation) to S2 (recombined - ionized electrons) depends on the interaction/recoil type (in its charge)

- Electromagnetic interaction releasing electrons in the liquid → S2
- Disclaimer: NOT a single-body problem, it's a (complicated) 'stopping power' problem, so there is a (smaller) residual energy transferred into prompt scintillation → S1



S2 >> S1
Predominant S2 signal

Details about Nuclear Recoils



- Scattering off the nucleus, releasing prompt photons → S1
- Recall, NOT a single-body problem, so there is a fraction of energy transferred into *electronic stopping* which causes ionization → S2
- Expect a much weaker S2
- The response can be modelled with various approaches, which tend to be energy, material, and particle type dependent: Lindhard theory, Doke-Birk's saturation, scaling rules, etc

S1 ~ S2*

WIMPs (signal) or Neutrons (background)

Photons

- **Actually S2 is still higher but is significantly smaller**
- **Smaller enough to be a discriminator!**