Selecting LS4GAN Domains

Brett Viren

August 18, 2021

Topics

- Attempt to define jargon and notation.
- LS4GAN concept revisited with jargon and notation.
- Tour of and selecting **domains** for LS4GAN input.
- Production processing estimates.

Jargon and Notation

- Try to define terms that both mathematicians, computer scientists and physicists can understand.
- Inspiration taken from the DL book
  - Deep Learning, Goodfellow, Bengio and Courville.
- May need some iteration to get right.

Jargon

"event"	a **vector** of scalar numbers.
"domain"	the **set** of **possible events** over which a functional **transform** may be defined.
"tier"	a specific **domain** produced by a **given transform** applied over a **given domain**.
"dataset"	a subset of **specific events** from a **domain**

Comments on Jargon

- An **event** vector typically possesses an implied structure and semantic interpretation which are associated to a **domain**.
  - "A readout of 12 bit ADCs from one detector plane are organized in time-major 2D arrays of shape (n_{chan}, n_{tick})."
- A **transform** may be applied to a larger **domain** that contains a desired **domain** subset.
  - "We may apply signal processing to any 2D array but that is subject to garbage-in-garbage-out."
- A **tier** is defined by both its input **domain** and the **transform**.
  - Signal processing on simulation is a different tier than signal processing on detector data.
- A **dataset** carries implied **provenance** of the **events** from the initial **domain** and all **transforms** leading to its production.
  - Different **domains** are always distinct though we strive to find **transforms** that make their **datasets** indistinguishable.

In any case, it is easy for meaning of jargon to slip around....

Notation proposal

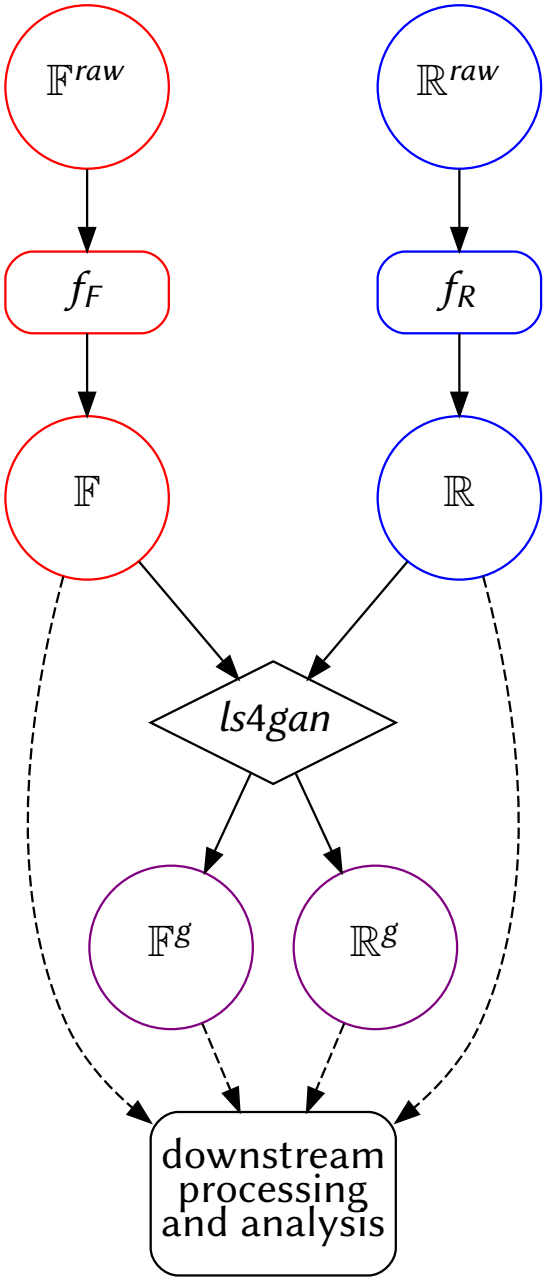
- $*$ a convolution
- s a scalar value
- $\mathbf{v}$ a vector or 1D array value (`\mathbf`)
- $\mathbf{A}$ a matrix or 2D array value (in Fourier interval space) (`\mathbf`)
- $\mathcal{A}$ a matrix or 2D array value in Fourier frequency space (`\mathcal`)
- $\mathbb{D}$ a domain (`\mathbb`)
  - special case: $\mathbb{E}()$ is "expectation function"
- f a functional transform of events in some domain
- $\mathbb{D}^f$ a tier domain qualified by f applied to domain $\mathbb{D}$
- $\mathbf{d}$ a possible event from domain $\mathbb{D}$ (`\mathbf`)
  - one scalar element of event vector: $d_i = \mathbf{d}[i]$
- $\mathbb{D}_j$ one specific dataset j drawn from domain $\mathbb{D}$
- $\mathbf{d}_k$ one specific event k from dataset $\mathbb{D}_j$

LS4GAN concept revisited

Present again the LS4GAN concept following jargon and notation.

LS4GAN concept

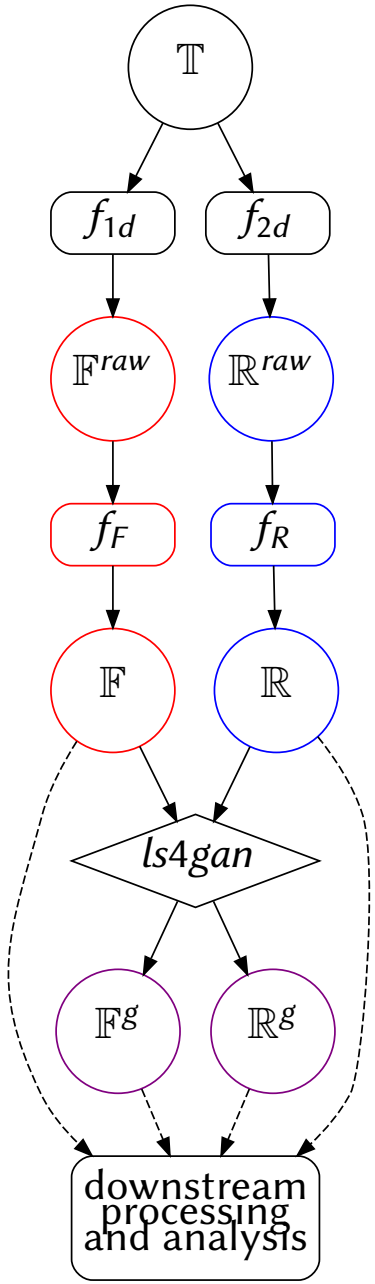
- Source of **datasets** from two **domains**
  - *eg*: $\mathbb{F}^{raw}$ tier is *fake* "raw" data from simulation,
  - and $\mathbb{D}^{raw}$ tier is *real* "raw" data from detector.
- Apply some **transform** f_X to each
  - may have $f_F \approx f_R$, *eg*, as with signal processing.
- Defines two data **tiers** for input to **LS4GAN**
  - Train LS4GAN to map $\mathbb{F} \leftrightarrow \mathbb{R}$
  - Apply trained LS4GAN generator: $ls4gan(\mathbb{F}; \mathbb{R}) \rightarrow \mathbb{F}^g$
  - Such that $|\mathbf{f}_i^g - \mathbf{f}_i|$ and $|\mathbb{E}(\mathbf{f}^g) - \mathbb{E}(\mathbf{r})|$ are small.
- Exploit approximately-matched **event pairs**
  - $\mathbf{f}^g$ is "corrected simulation" corresponding to $\mathbf{f}$
  - Pairs propagate simulation systematic uncertainty: $|f_X(\mathbf{f}^g) - f_X(\mathbf{f})|$ or $|f_X(\mathbf{r}^g) - f_X(\mathbf{r})|$.



LS4GAN sim/sim special case

Special case: both *fake* and *real* datasets are simulated.

- Both **datasets** are from **tiers** derived from common *ground truth* **domain** $\mathbb{T}$.
- Gain a new type of matched pair: $(\mathbf{f}_i, \mathbf{r}_i)$
  - Represents a "perfect" LS4GAN mapping
- Enables special case event-by-event validation:
  - $\mathbf{f}_i^g \sim \mathbf{r}_i$
  - $\mathbf{r}_i^g \sim \mathbf{f}_i$



LS4GAN Input Domain Requirements

Requirements for choice of *real* and *fake* domains:

- Minimize complexity of what LS4GAN must learn.
  - Choose $\mathbb{F}$ and $\mathbb{R}$ to be "nearly overlapping".
  - Exhaust our domain expertise to remove unimportant features.
  - Want $|\mathbb{E}(\mathbf{f}) - \mathbb{E}(\mathbf{r})|$ to already be small.
- Maximize efficiency to keep and minimize bias change physics information.
  - Best to operate close to "low level, raw data features".
  - Do not assert a model which changes measured physical quantities.
    - ★ LS4GAN systematics strategy should uncover blunders here.
- Generated domains $\mathbb{F}^g$ and $\mathbb{R}^g$ must contain events that "look like normal data" to downstream analysis.

Some LArTPC Data Tiers

$\mathbb{K}^0$ initial, pre-interaction kinematics: $\{\vec{p}, \vec{r}, m\}$.

$\mathbb{K}^f$ final state, post-interaction kinematics: $\sigma, \{\vec{p}, \vec{r}, m\}$.

$\mathbb{P}$ particle tracks/trajectories: $\{E(t), \vec{r}(t)\}$.

$\mathbb{D}^0$ initial ionization electron distribution: $\{\vec{r}, q, t\}$.

$\mathbb{D}$ post-drift ionization electron distribution: $\{\vec{r}, q, t, \vec{\sigma}\}$.

$\mathbb{S}^i$ "true signal": per-plane, time-sliced, per-impact i, tomographic view of D: array size $\mathcal{O}(10k \times 10k)$.

"Observability boundary". Above is fine-grained detailed information we may only simulate or estimate/reconstruct.

$\mathbb{A}$ time *vs* electronics channel ADC waveforms, $S * R_\alpha$: $\mathcal{O}(1k \times 10k)$.

$\tilde{\mathbb{S}}$ estimate channel-average signal $\overline{S}$: $\mathcal{O}(1k \times 10k)$.

$\mathbb{D}_{wc}$ estimate D, *eg* Wire-Cell tomographic reconstruction from $\overline{S}_\alpha$.

$\mathbb{D}_{fm}$ estimate D^0, *eg* WC charge/light matching for T_0 of clusters of D_{wc}.

... further reconstruction toward P, K^f and K^0

Some LArTPC Data Tiers

$\mathbb{K}^0$ initial, pre-interaction kinematics: $\{\vec{p}, \vec{r}, m\}$.

$\mathbb{K}^f$ final state, post-interaction kinematics: $\sigma, \{\vec{p}, \vec{r}, m\}$.

$\mathbb{P}$ particle tracks/trajectories: $\{E(t), \vec{r}(t)\}$.

$\mathbb{D}^0$ initial ionization electron distribution: $\{\vec{r}, q, t\}$.

$\mathbb{D}$ post-drift ionization electron distribution: $\{\vec{r}, q, t, \vec{\sigma}\}$.

$\mathbb{S}^i$ "true signal": per-plane, time-sliced, per-impact i, tomographic view of D: array size $\mathcal{O}(10k \times 10k)$.

"Observability boundary". Above is fine-grained detailed information we may only simulate or estimate/reconstruct.

$\mathbb{A}$ **time *vs* electronics channel ADC waveforms, $S * R_\alpha$: $\mathcal{O}(1k \times 10k)$.**

$\tilde{\mathbb{S}}$ **estimate channel-average signal $\overline{S}$: $\mathcal{O}(1k \times 10k)$.**

$\mathbb{D}_{wc}$ estimate D, *eg* Wire-Cell tomographic reconstruction from $\overline{S}_\alpha$.

$\mathbb{D}_{fm}$ estimate D^0, *eg* WC charge/light matching for T_0 of clusters of D_{wc}.

... further reconstruction toward P, K^f and K^0

$\mathbb{A}$: ADC vs $\tilde{\mathbb{S}}$: Signal domains

Next, describe these two domains, how they are produced and how they may be considered against the requirements.
Optional slides:

- Aside into special case of sim+sigproc using matched responses
- Look at “splat” as an approximation to signals.

ADC Waveforms

Two family of ADC tiers to consider:

- $\mathbb{A}^d$ as produced by the **detector** and its electronics.
  - May be considered a family of tiers parameterized by detector conditions and DAQ trigger.
- $\mathbb{A}^s$ as produced by **simulation** of particle interaction physics, detector physics and electronics operation.
  - May be considered a family of tiers with same parameters as $\mathbb{A}^d$ in addition to model parameters (physics and detector response).

Next, production of $\mathbb{A}^s$ is described.

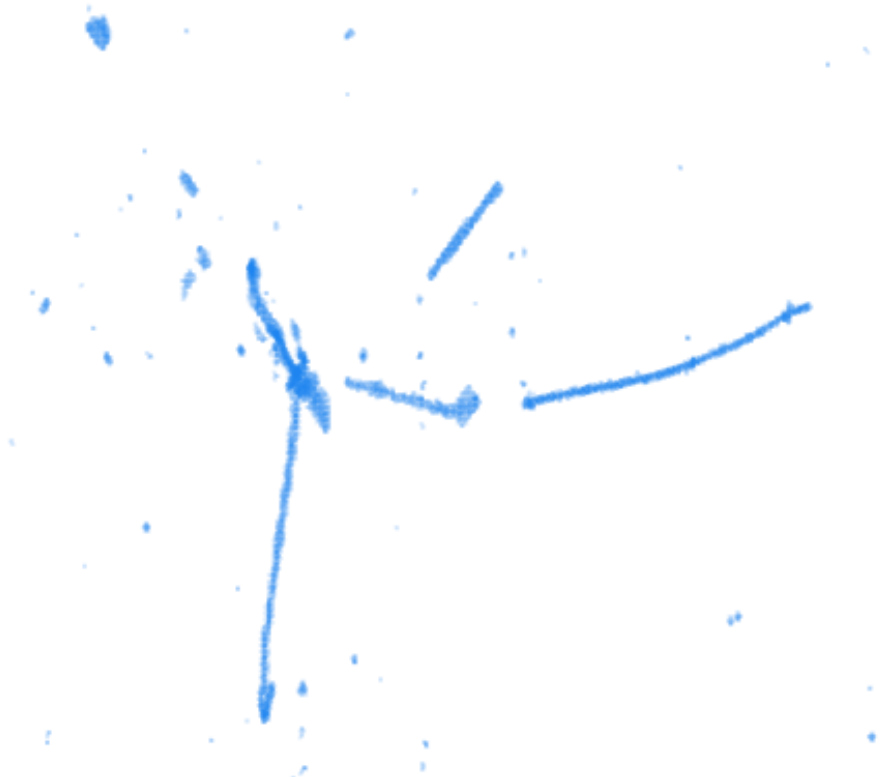
$$\text{interaction} \rightarrow \text{tracking} \rightarrow \text{drift} \rightarrow \text{projection} \rightarrow \text{convolution} \rightarrow \text{digitization}$$

Simulation - interaction + tracking

Interaction and tracking

- initial particle kinematics from various packages.
- particle interaction and nuclear from, eg GENIE, Geant4 or FLUKA
- particle tracking by Geant4

Produces D^0, initial ionization electron distribution.



Example event displayed by BNL/EDG's "Bee" 3D interactive web-based visualization system.
https://www.phy.bnl.gov/wire-cell/bee/set/23/event/0/

Simulation - drift + projection

Drift:

- ionization electrons at a point for each tracking step transported to near the wire planes
- electron absorption, diffusion, total charge fluctuation
- results in a truncated 2D Gaussian distribution for each.

Project:

- Gaussian distributions binned to a 2D **wire pitch** *vs* **drift step** coordinate system associated with each wire plane.
- pitch bins are $\frac{1}{10}^{th}$ pitch wide

Produces S^i, the 2D ionization charge distribution for each relative impact bin i for a given wire/strip plane.

Simulation - convolution

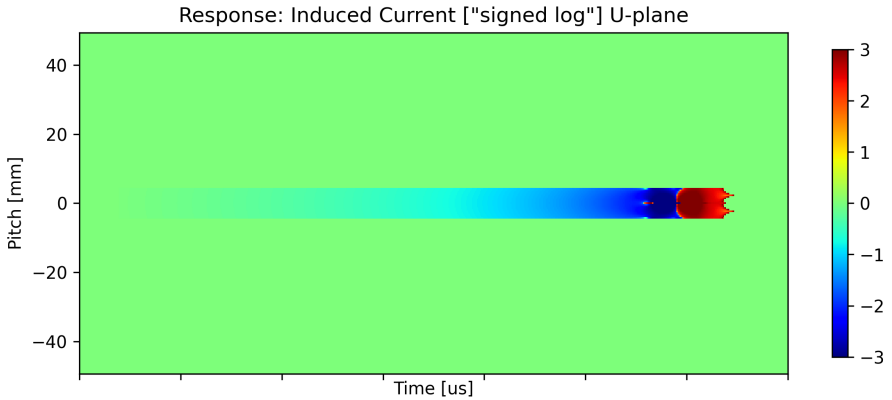
The detector response is applied to produce a "pre-measure" as an "interlaced-convolution" plus noise:

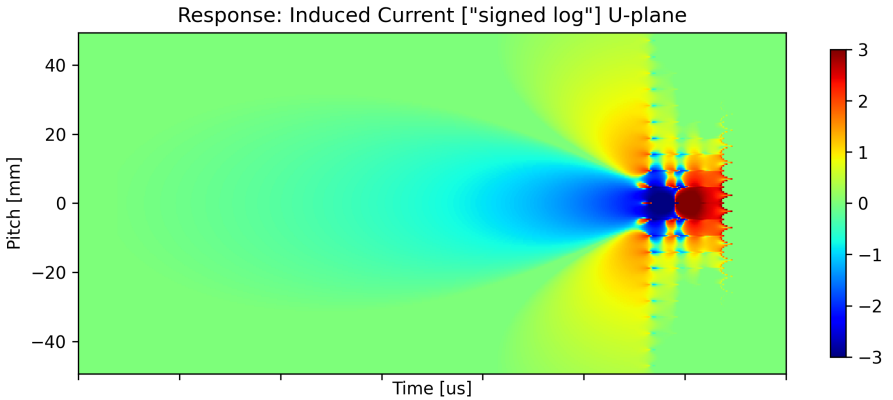
$$M_v = \sum_i R^i * S^i + N$$

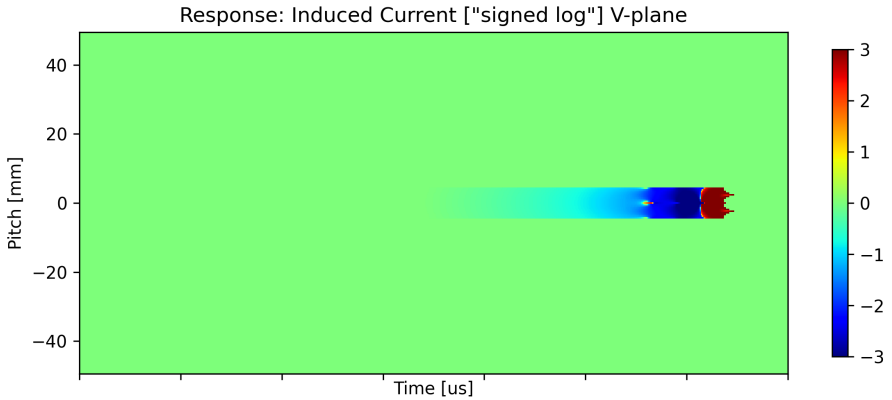
- $S^i \sim (n_{wire}, n_{step})$ array giving counts of ionization electrons in each *impact bin i* counted relative to the nearest wire.
- $R^i \sim (n_{wire'}, n_{step'})$ array giving induced *response* from one electron falling on impact bin i relative to given wire and over one drift step.
- $N \sim (n_{wire}, n_{step})$ an array of noise[1]
- $M_v \sim (n_{wire}, n_{tick})$ represents a contribution to voltage from given wire during a given sampling period (tick).
- $\sum_i$ is over typically $n = 10$ impact bins per wire.

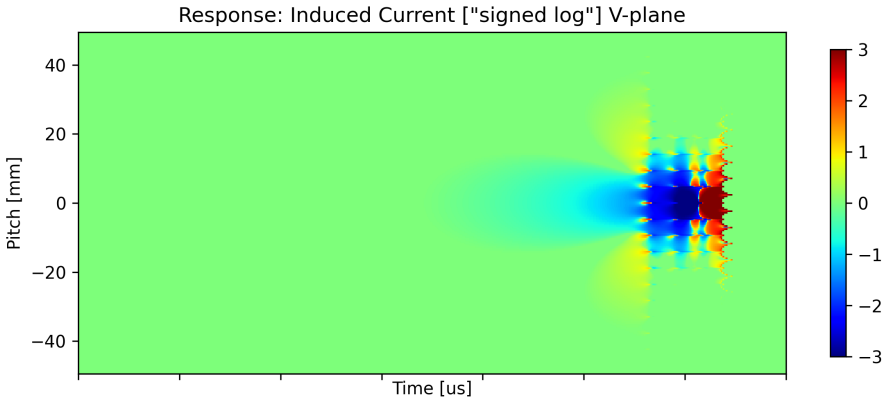
[1] We do not write it explicitly but N is subject to electronics response which is also forms part of R^i.

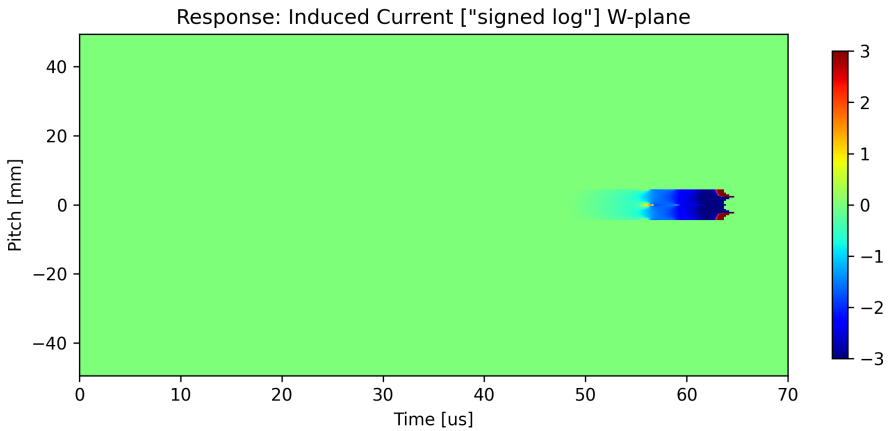
Example R^i convolution kernels: q1D and 2D

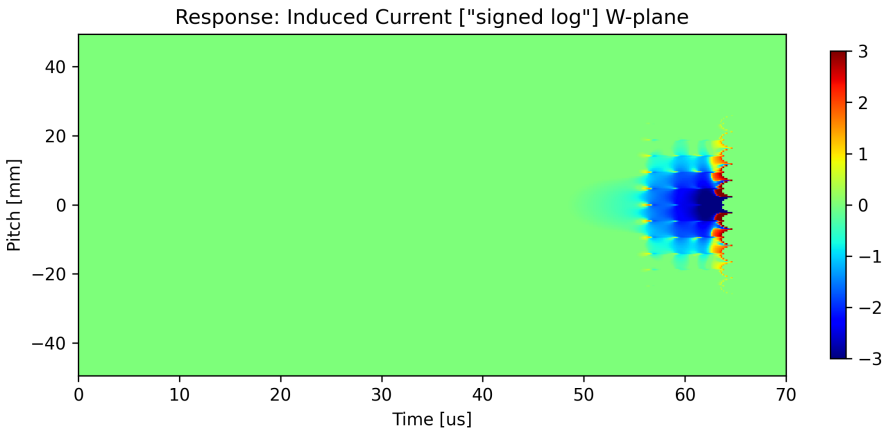












Simulation - digitization stage

And the final ADC measure:

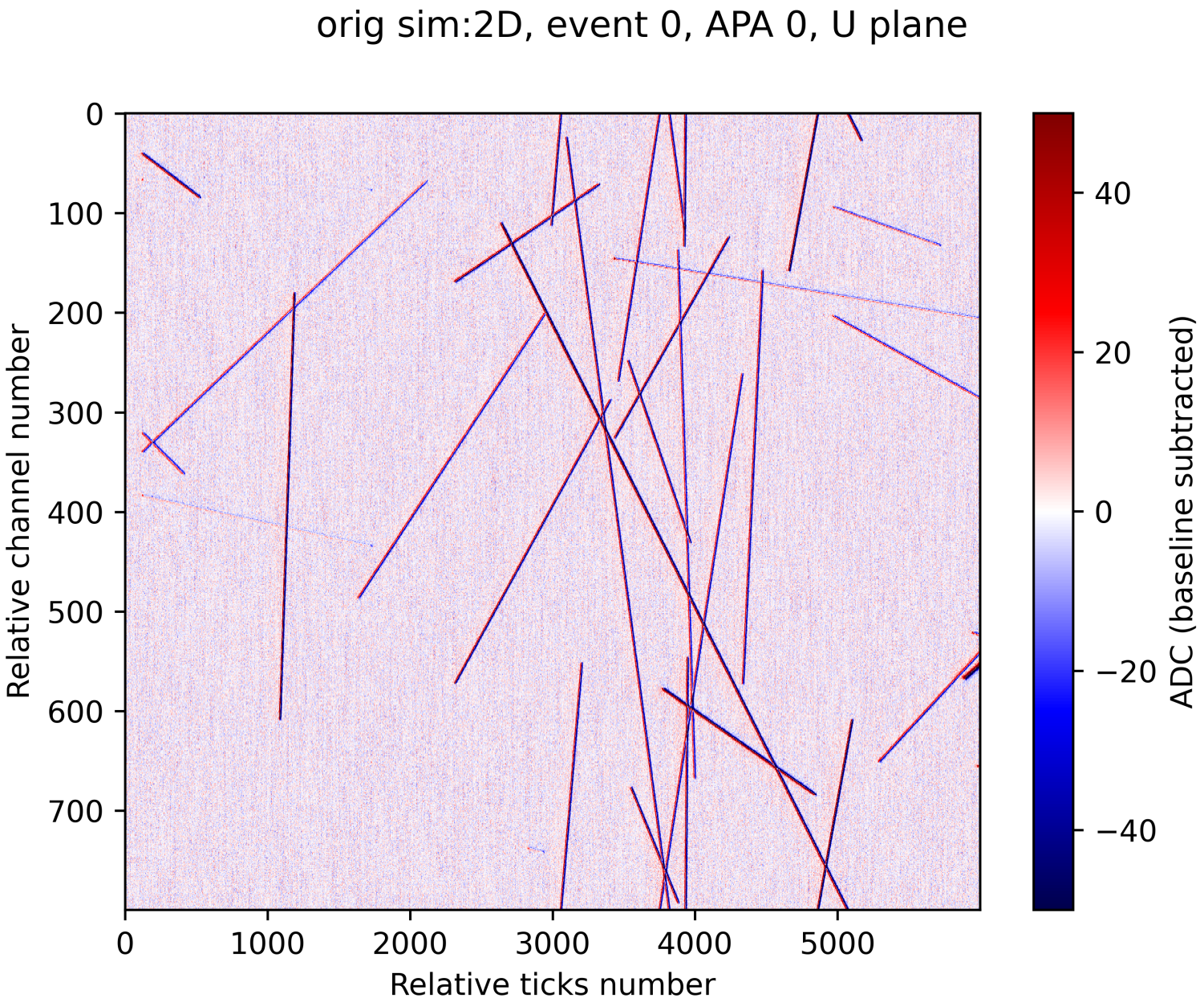
$$M = floor(s \cdot (o + CM_v))$$

- s a scale factor (ADC/Volt ratio)
- o an offset (baseline)
- $C \sim (n_{chan}, n_{wire})$ channel map (sums some wire/rows of M_v)
- $M \sim (n_{chan}, n_{tick})$ ADC count on given channel at given sampling period (tick).[2]
- $floor()$ integer-truncation function

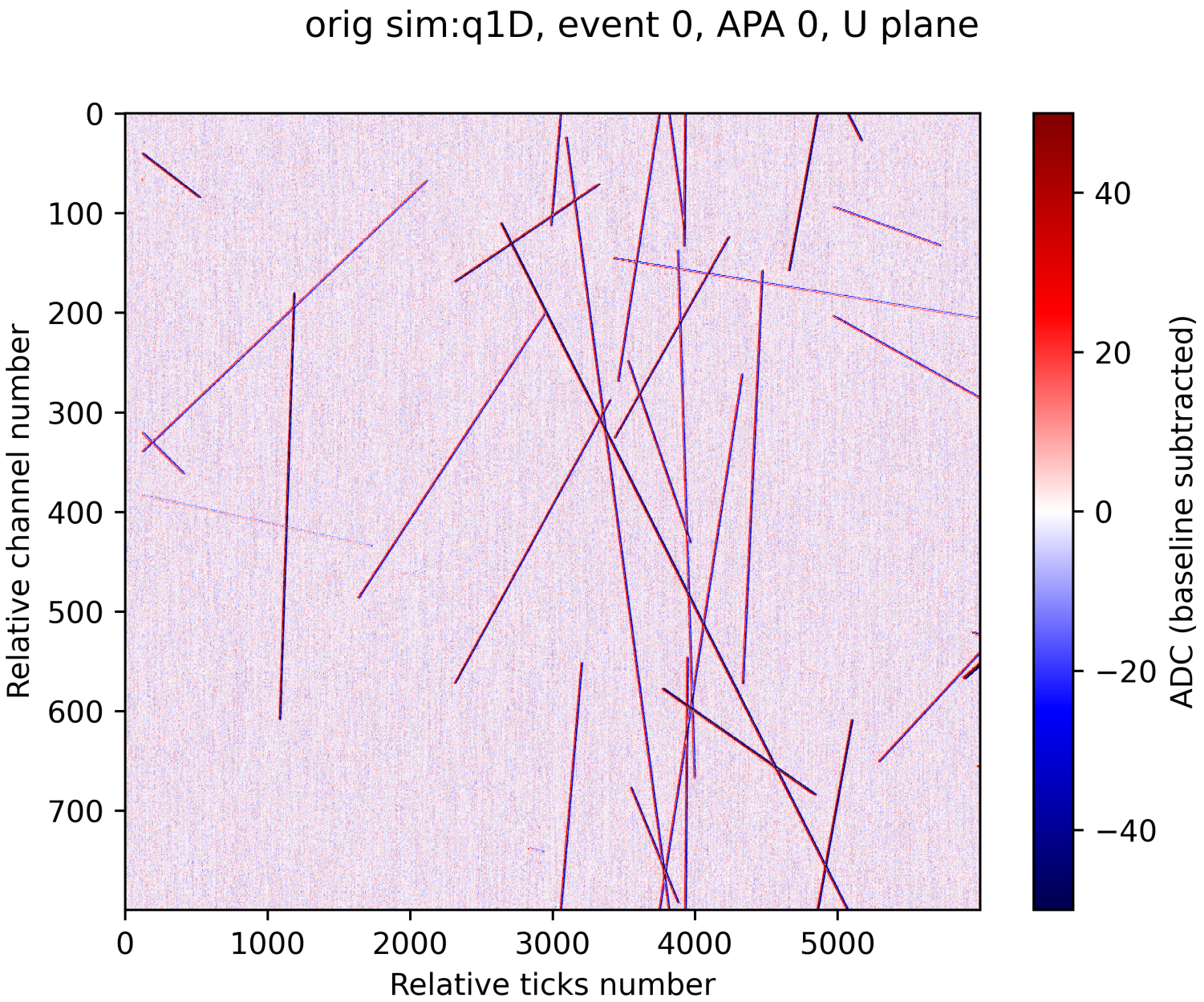
For $C = \mathbb{I}$ (no wrapped wires) we may approximate $M_v \to M \approx M_v$ up to the overall scaling.

[2] For convenience we choose *tick* and *step* to identically.

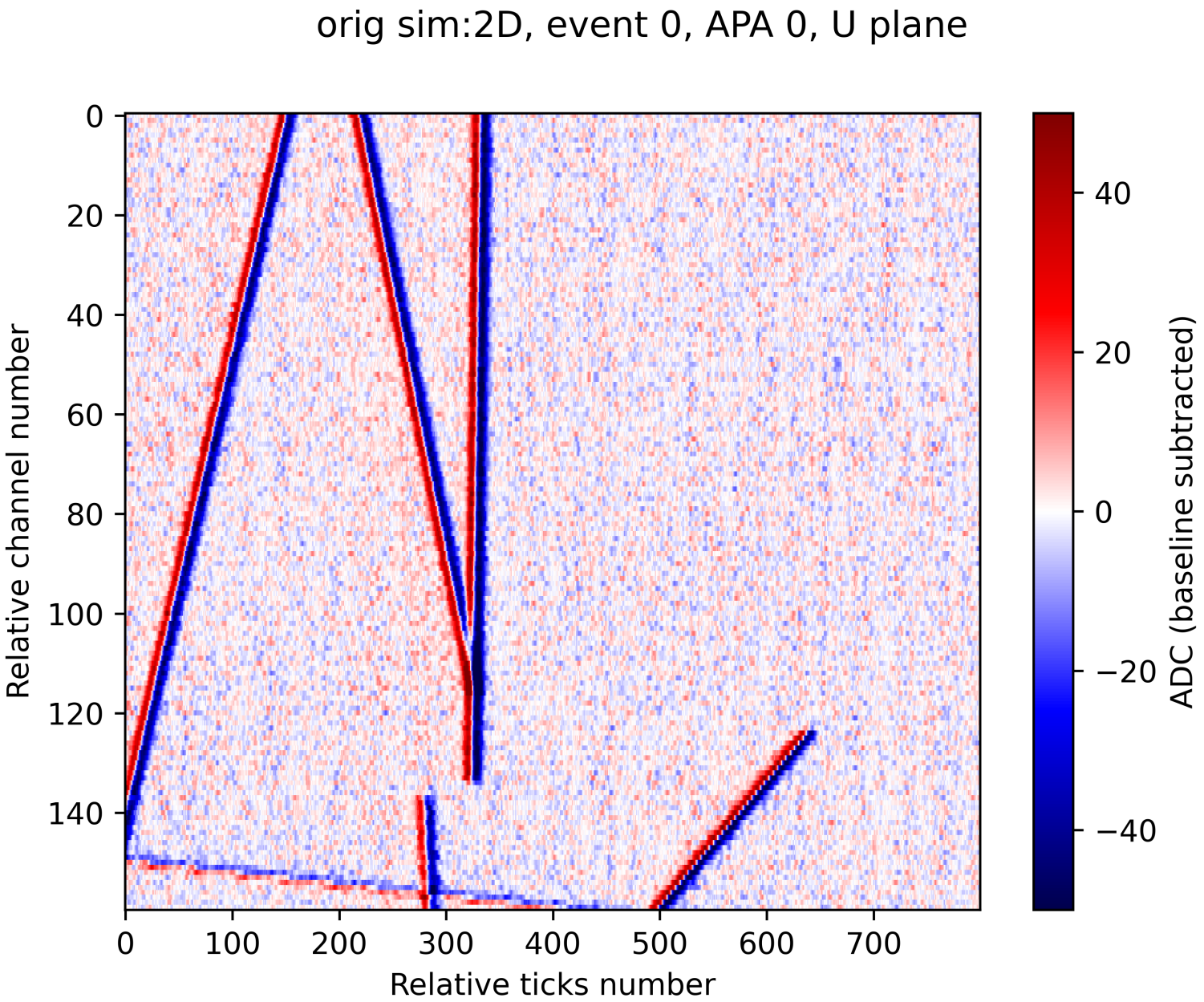
ADC 2D response



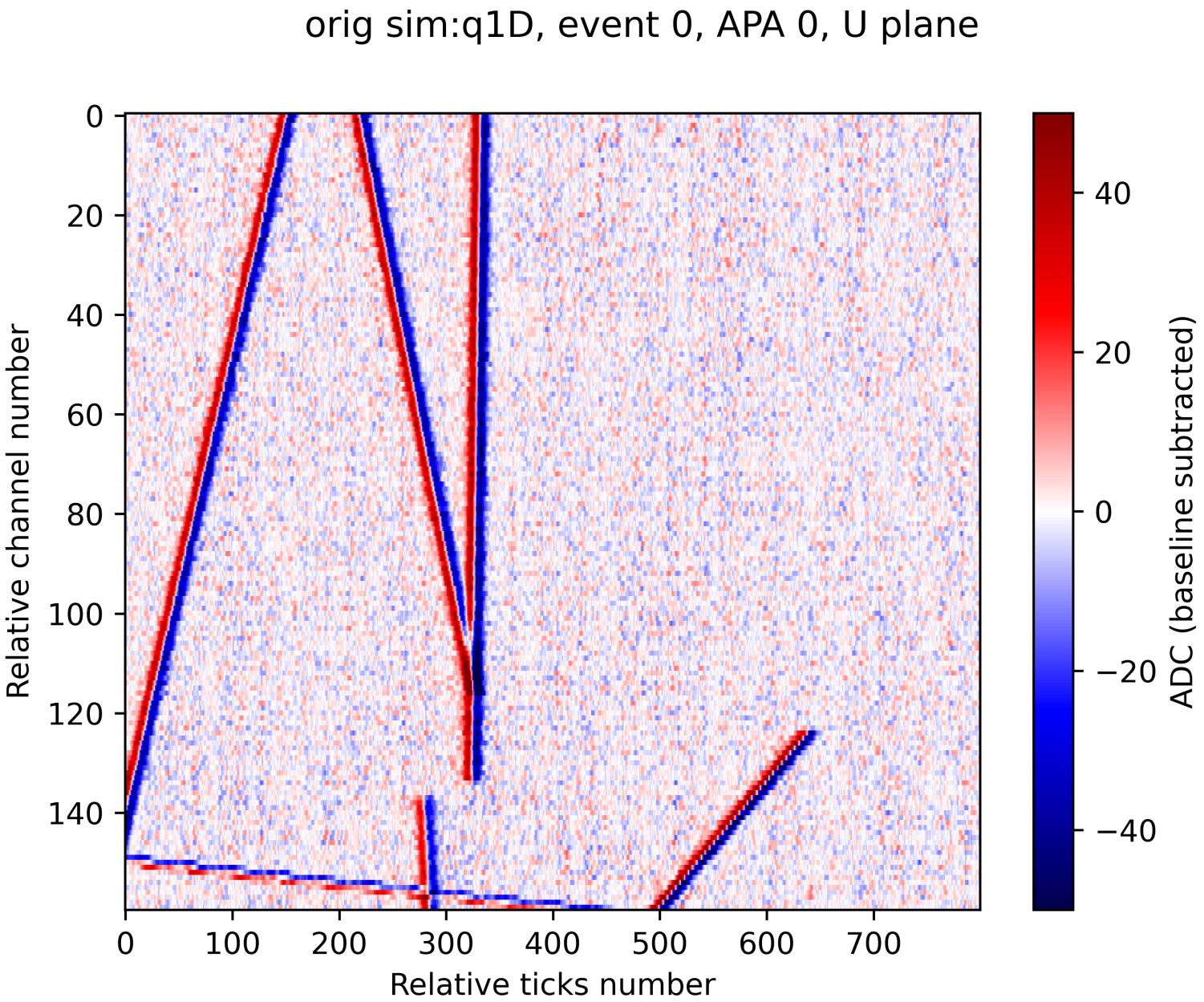
ADC q1D response



ADC 2D response, zoom



ADC q1D response, zoom



LS4GAN with ADC tiers

Problems in selecting ADC for LS4GAN's $\mathbb{F}/\mathbb{R}$ domains

- LS4GAN must learn **first-order differences** between $\{R^i\}$ and R^{nature}
  - The **bipolar induction-plane signals** are "faint and fragile".
  - Different R^i gives very different features.
- Simulation fails to reproduce detector's "rich" coherent noise features which LS4GAN would need to learn.
  - We can model this better but also know how to (mostly) remove it in "noise filtering" prior to signal processing.

Signal Processing

The signal tiers are made produced by optional noise-filtering and signal processing which consists of two main parts:

- A deconvolution of the ADC waveforms with an average response
- Identifying regions-of-interest (signal-ROI) which likely contain signal.

Skip noise filtering (interesting on its own) and give some formalism on signal processing.

Transform ADC to Signal Waveforms

Define an *average response* over the sub-pitch impact bins:

$$R \equiv \frac{1}{n} \sum_i R^i$$

and an *average true signal* $S \sim (n_{wire}, n_{step})$ such that

$$R * S \equiv \sum_i R^i * S^i$$

to allow rewriting the measure as:

$$M = R * S + N$$

and an **estimated average pre-signal** is recovered:

$$\tilde{S}' = R^{-1} * M$$

Signal Regions Of Interest

Again pre-signal:

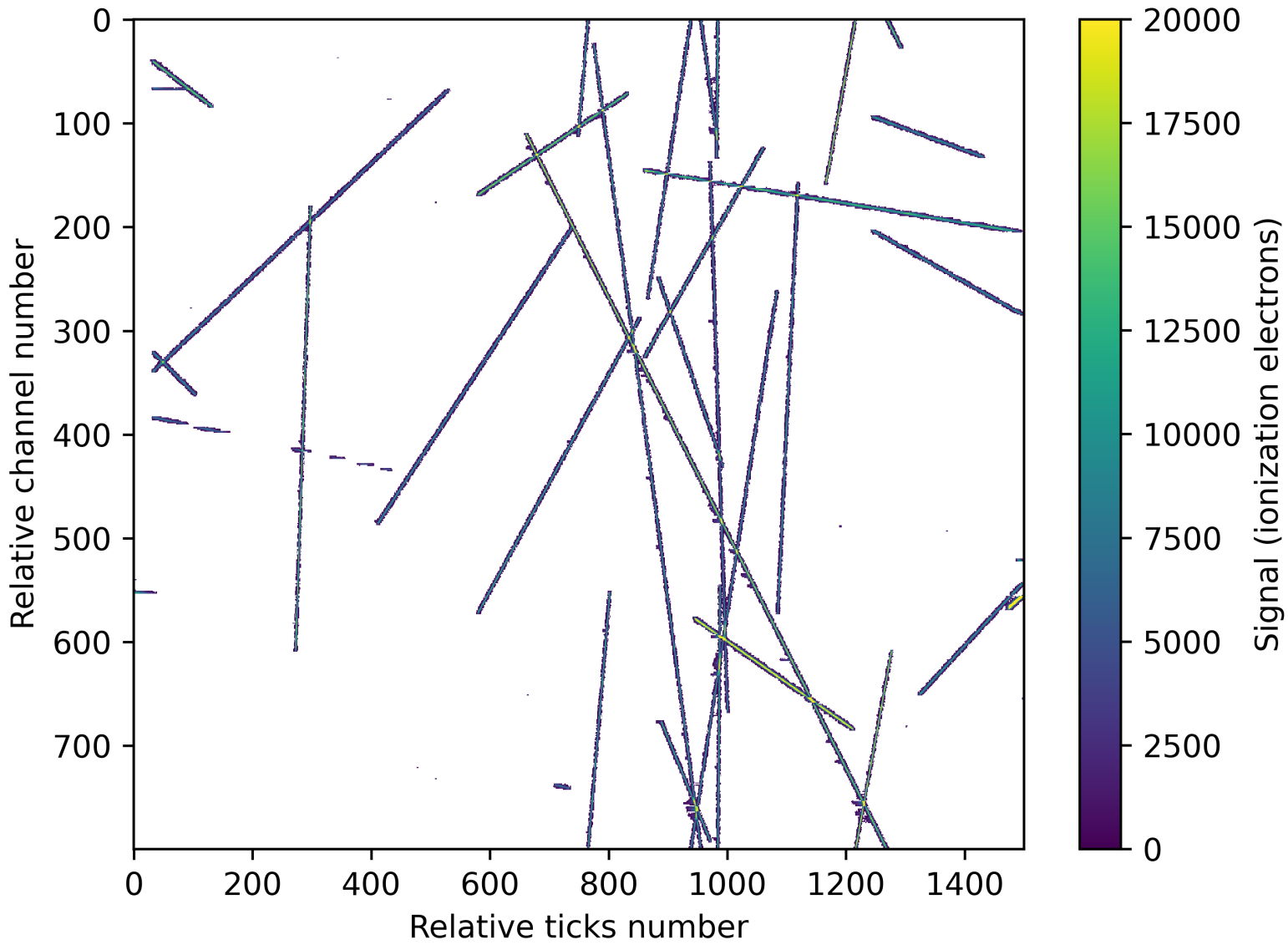
$$\tilde{S}' = R^{-1} * (\sum_i R^i * S^i + N)$$

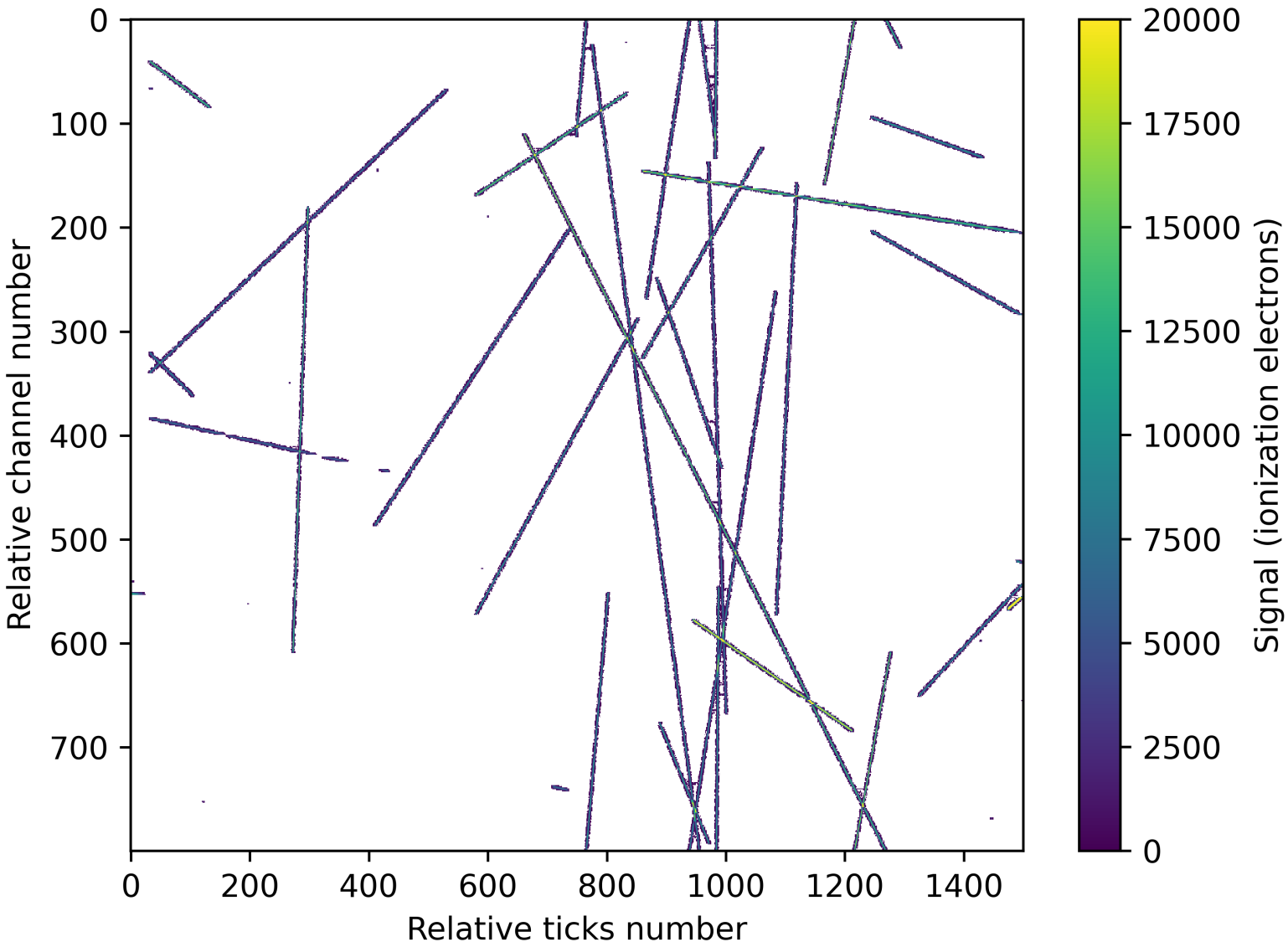
For the bipolar channels the $\mathcal{R}$ term tends to zero at zero frequency.

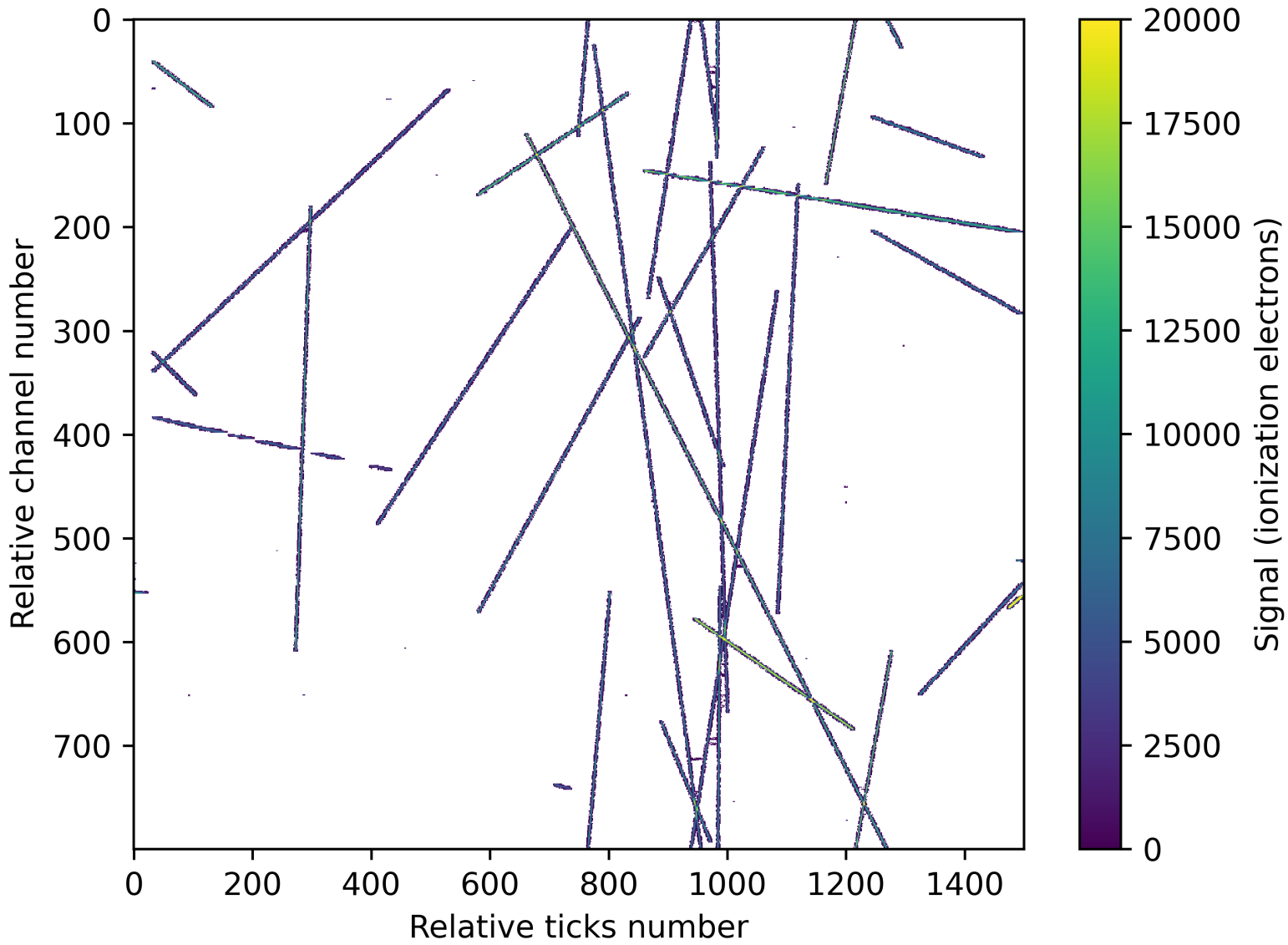
- Any residual **low-frequency noise is amplified**.
- We identify regions-of-interest containing signal in the time domain where small signal is above amplified low-frequency noise.

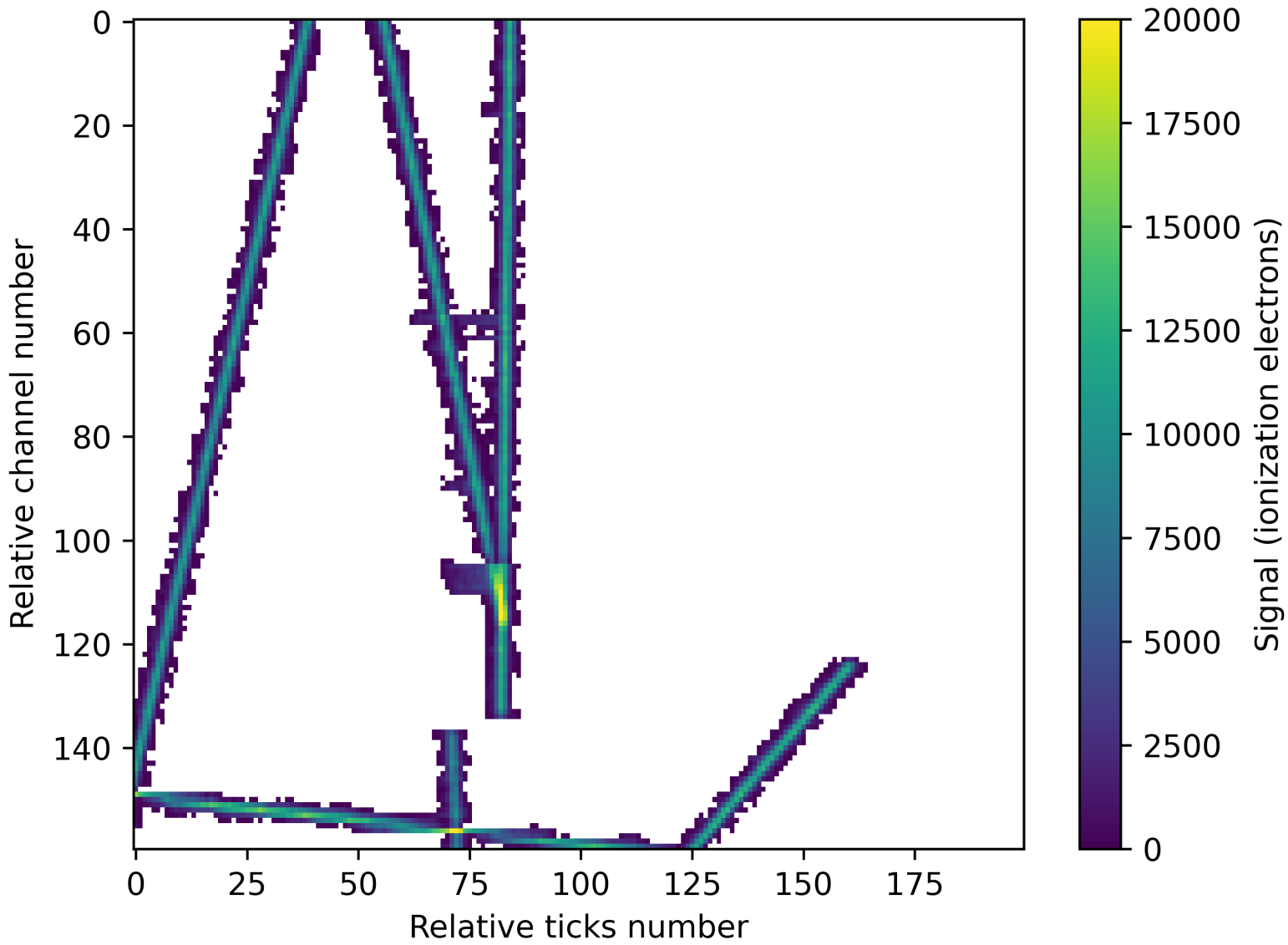
$$\tilde{S} = roi(\tilde{S}')$$

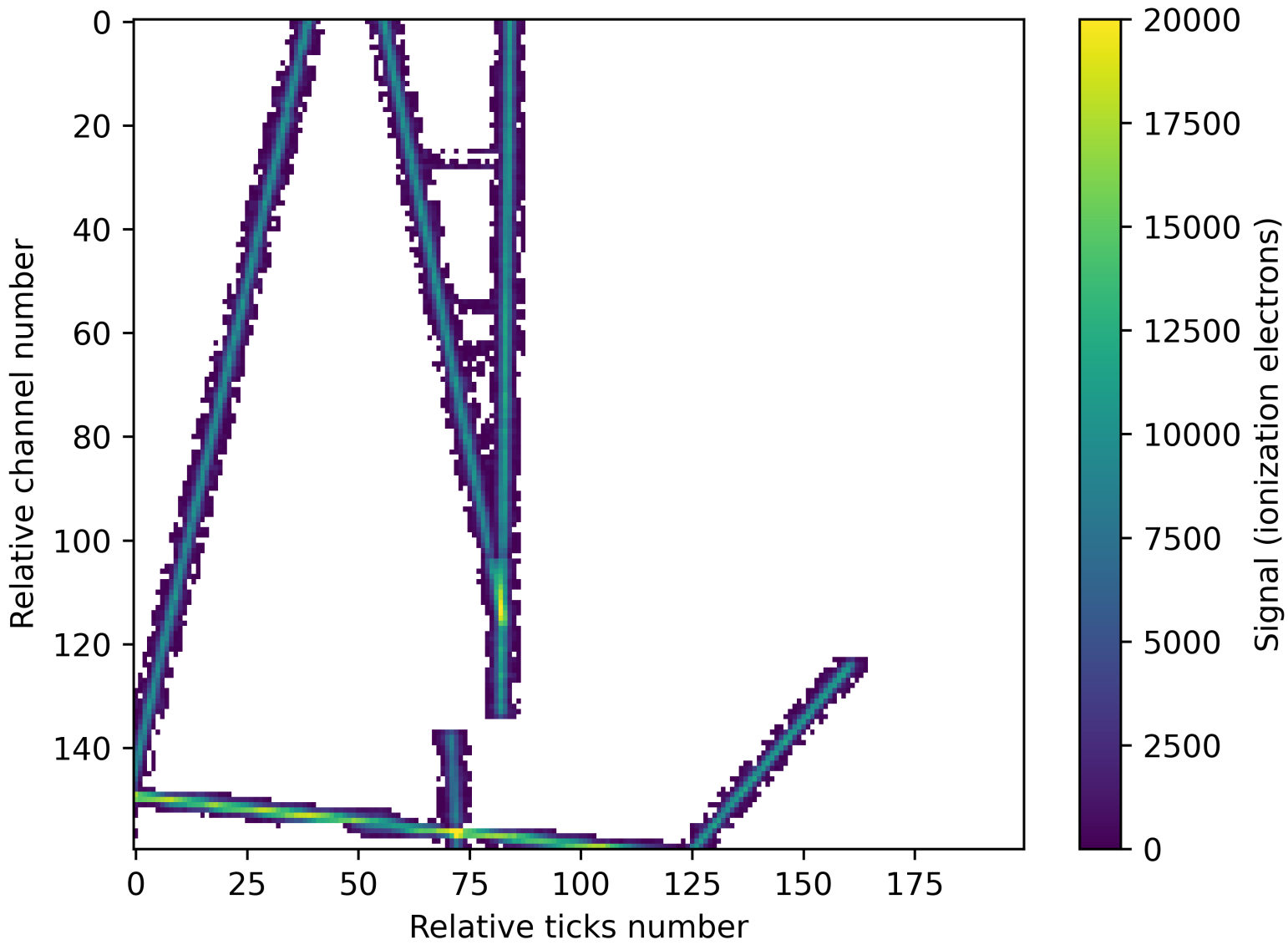
The *roi*() function is "highly heuristic" but can be thought of as an *adaptive high pass filter*.

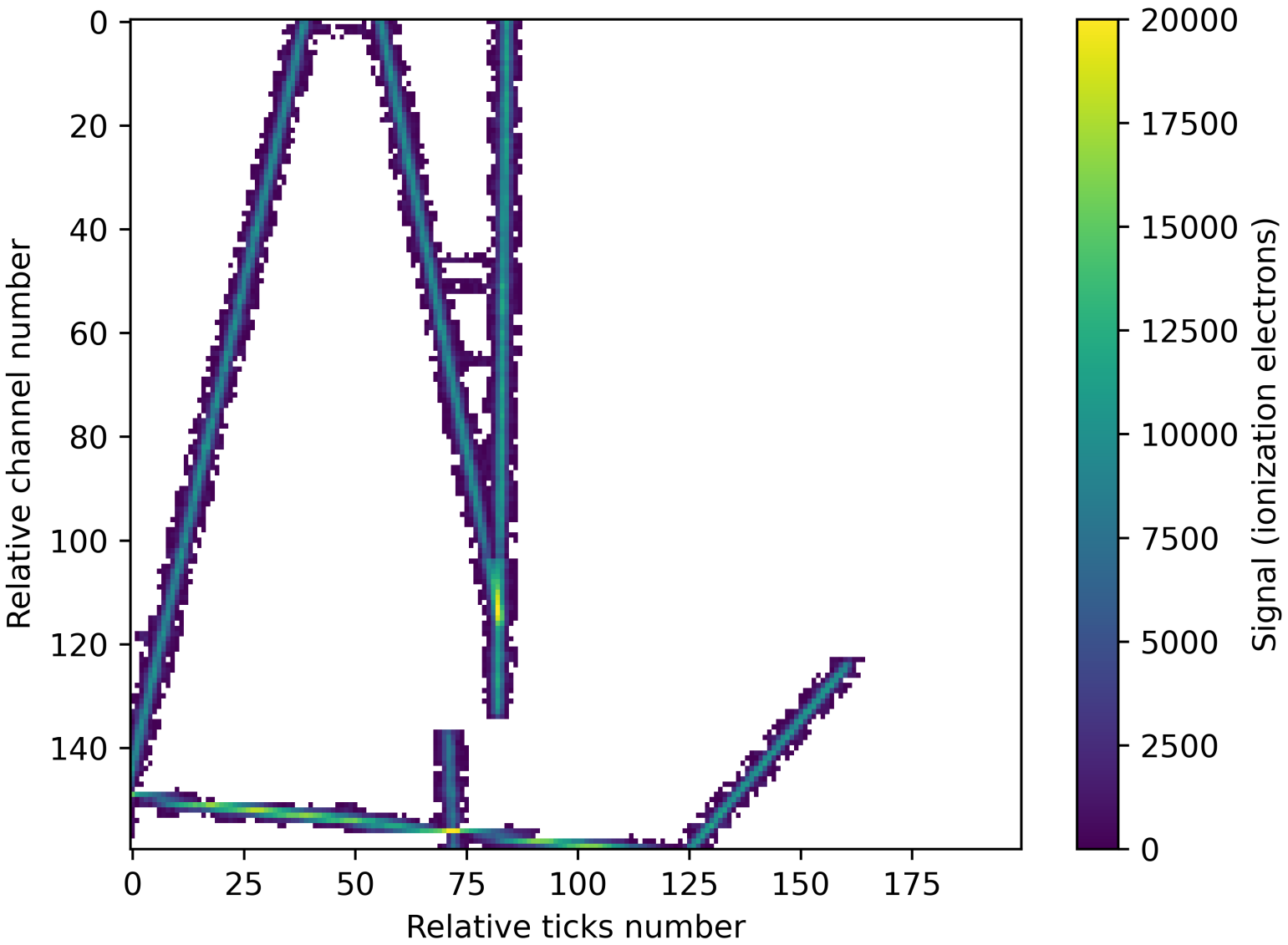

gauss sim:2D/SP:q1D, event 0, APA 0, U plane
Relative channel number
0
100
200
300
400
500
600
700
Relative ticks number
0
200
400
600
800
1000
1200
1400
Signal (ionization electrons)
20000
17500
15000
12500
10000
7500
5000
2500
0


gauss sim:q1D/SP:q1D, event 0, APA 0, U plane
Relative channel number
0
100
200
300
400
500
600
700
Relative ticks number
0
200
400
600
800
1000
1200
1400
Signal (ionization electrons)
20000
17500
15000
12500
10000
7500
5000
2500
0


gauss sim:2D/SP:2D, event 0, APA 0, U plane
Relative channel number
0
100
200
300
400
500
600
700
Relative ticks number
0
200
400
600
800
1000
1200
1400
Signal (ionization electrons)
20000
17500
15000
12500
10000
7500
5000
2500
0


gauss sim:2D/SP:q1D, event 0, APA 0, U plane
Relative channel number
0
20
40
60
80
100
120
140
Relative ticks number
0
25
50
75
100
125
150
175
Signal (ionization electrons)
20000
17500
15000
12500
10000
7500
5000
2500
0


gauss sim:q1D/SP:q1D, event 0, APA 0, U plane
Relative channel number
0
20
40
60
80
100
120
140
Relative ticks number
0
25
50
75
100
125
150
175
Signal (ionization electrons)
20000
17500
15000
12500
10000
7500
5000
2500
0


gauss sim:2D/SP:2D, event 0, APA 0, U plane
Relative channel number
0
20
40
60
80
100
120
140
Relative ticks number
0
25
50
75
100
125
150
175
Signal (ionization electrons)
20000
17500
15000
12500
10000
7500
5000
2500
0

Aside: a special case of matched R in simulation

Reminder

- Simulation uses individual $\{R^i\}$ for numeric convolution.
- Detector uses unknown and continuous $R^{nature}(\vec{r})$ for a "true convolution".
- Both are deconvolved with the same average $R \equiv \frac{1}{n}\sum_i R^i$ in signal processing.

A special case arises for **simulation**:

$$\tilde{S}' \approx \mathcal{F}^{-1}\left(\frac{\mathcal{F}(M)}{\mathcal{F}(R)}\right) \equiv \mathcal{F}^{-1}\left(\frac{\mathcal{M}}{\mathcal{R}}\right)$$

$$\tilde{S}' \approx \sum_i \mathcal{F}^{-1}\left(\frac{\mathcal{R}^i}{\mathcal{R}}\mathcal{S}^i\right) \equiv \sum_i \mathcal{F}^{-1}\left(\tilde{\mathcal{R}}^i\mathcal{S}^i\right) = \sum_i \tilde{R}^i * S^i$$

Where this new $\tilde{\mathcal{R}}^i$ term is the Fourier dual to a kind of *residual response* $\tilde{R}^i$ for the ith impact position.

Aside: what is the residual response $\tilde{R}^i$?

It represents the **relative deviation** of an impact response from the local wire-average response (in Fourier frequency space).

$$\mathcal{F}(\tilde{R}^i) \equiv \tilde{\mathcal{R}}^i = \frac{\mathcal{R}^i}{\sum_i \mathcal{R}^i}$$

- For non-trivial R (q1D, 2D) the cancellation is only approximate.
  - In 1D response and ignoring noise there is **perfect cancellation**.
- $\tilde{R}^i$ expected to be nearly zero everywhere except with small, high-frequency wiggles at steps very near to the wires.
- Effectively, a smearing of S^i is produced.
  - Potentially we may be able to replace $\sum_i \tilde{R}^i * S^i \to G * S$, where G is Gaussian smear.
  - This would give a further $n = 10\times$ speed up of sim+sigproc.
- Nature of the smearing will itself vary between detectors
  - Larger smearing expected using strips+holes than wires.

WCT's `DepoSplat` component provides this approximate sim+sigproc.

Aside: Noise formalism

We dropped noise for simplicity. Adding it back now:

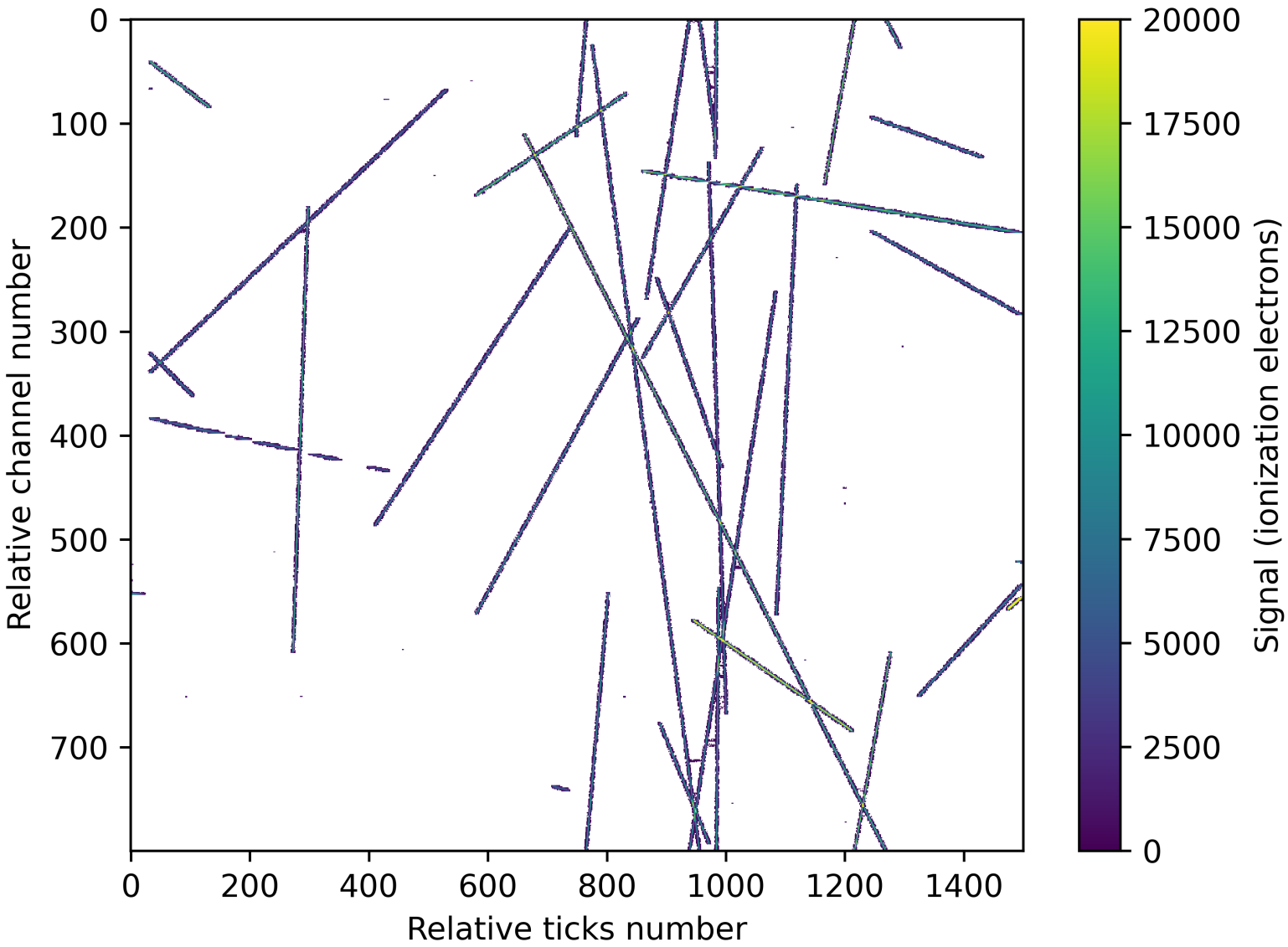
$$\tilde{S} \approx \sum_i \mathcal{F}^{-1}\left(\frac{\mathcal{R}^i\mathcal{S}^i + \mathcal{N}}{\mathcal{R}}\right) \equiv \sum_i \mathcal{F}^{-1}\left(\tilde{\mathcal{R}}^i\mathcal{S}^i + \tilde{\mathcal{N}}\right)$$

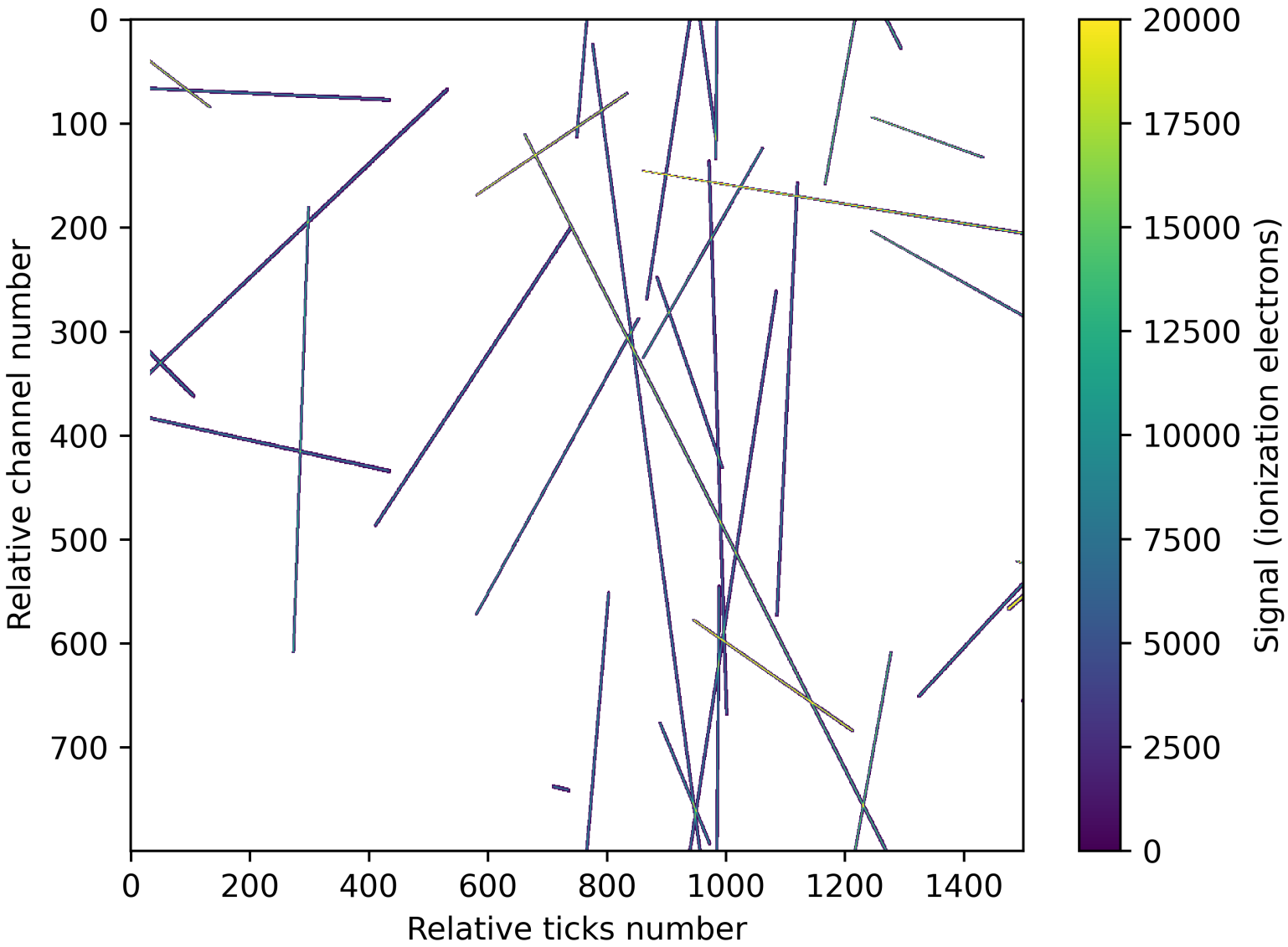
$$\mathcal{N} \equiv \mathcal{F}(N);\ \tilde{\mathcal{N}} \equiv \frac{\mathcal{N}}{\mathcal{R}}$$

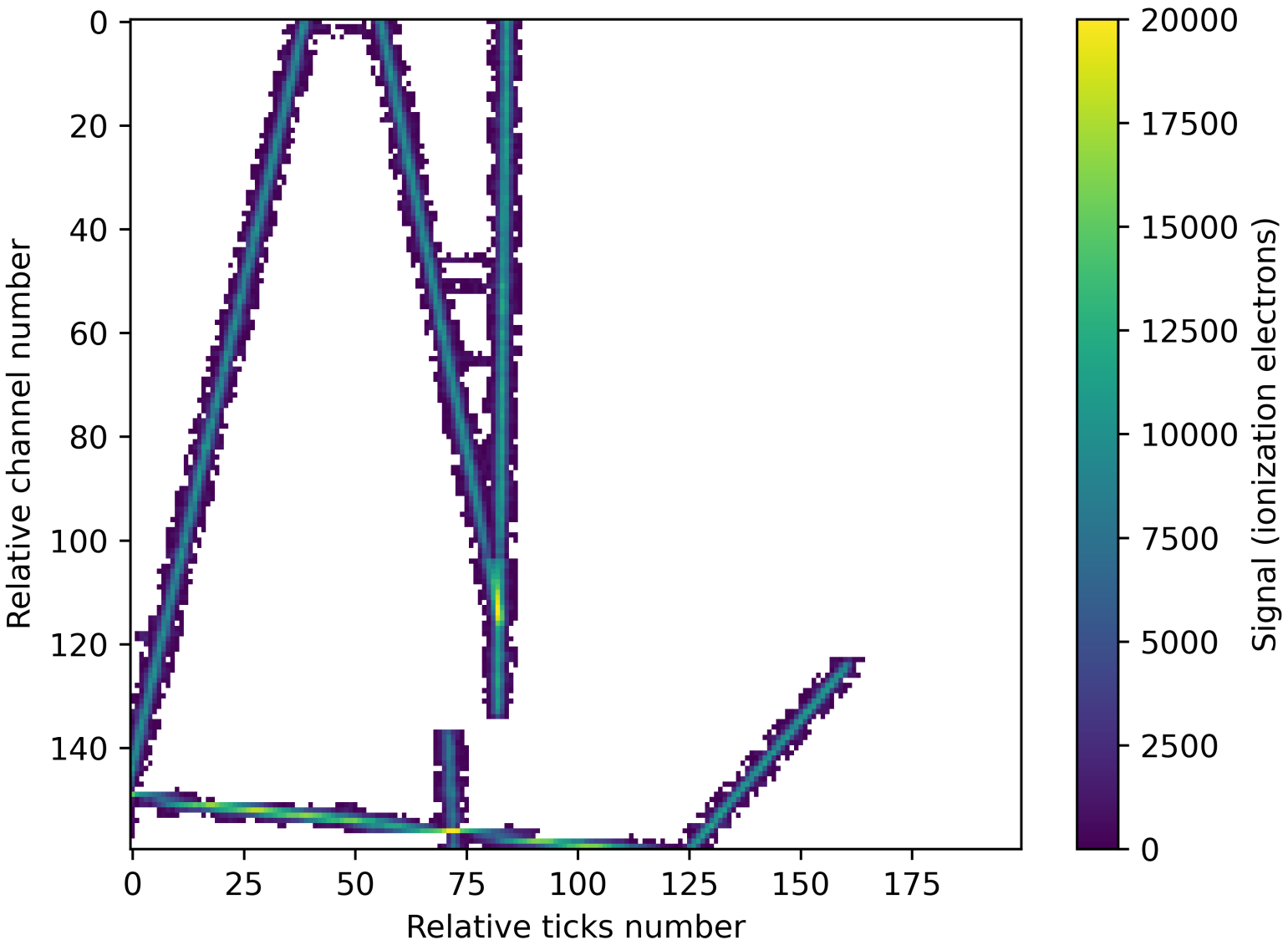
Thus the noise[3] in signals is seen as being "filtered" by the inverse of the average response.

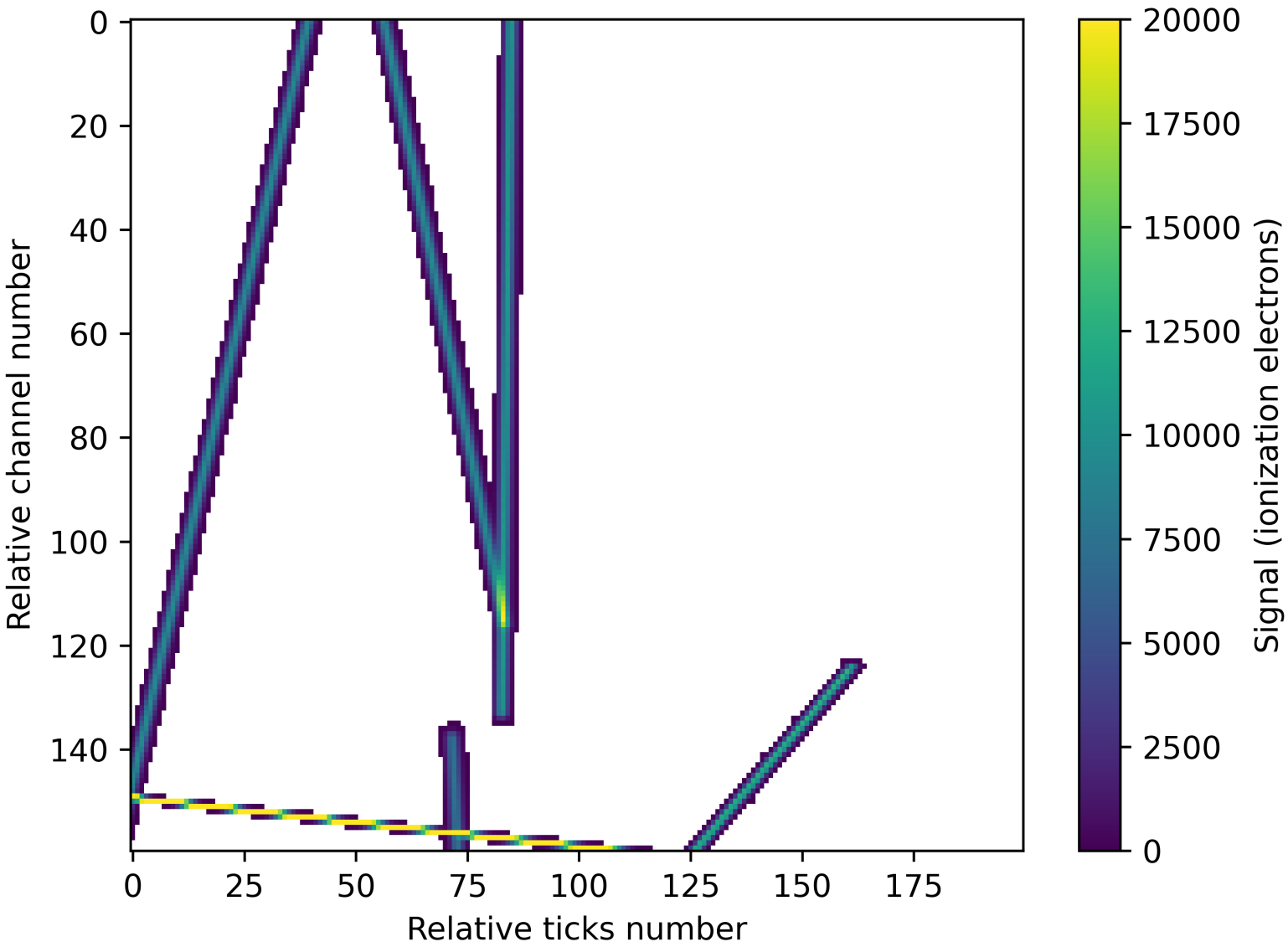
- We might generate $\mathcal{N}$ directly in Fourier space to save an FFT?

[3] Expanding on a footnote above, $\mathcal{R}^i = \mathcal{R}^i_{FR}\mathcal{R}_{ER}$ and really $\mathcal{N} = \mathcal{R}_{ER}\mathcal{N}'$ where $\mathcal{N}'$ is some "unshaped" noise from wire capacitance and resistance before the shaper circuit. $\mathcal{R}_{ER}$ thus cancels from the noise term but this seems not particularly exploitable.


gauss sim:2D/SP:2D, event 0, APA 0, U plane
Relative channel number
0
100
200
300
400
500
600
700
Relative ticks number
0
200
400
600
800
1000
1200
1400
Signal (ionization electrons)
20000
17500
15000
12500
10000
7500
5000
2500
0


splat splat, event 0, APA 0, U plane
Relative channel number
0
100
200
300
400
500
600
700
Relative ticks number
0
200
400
600
800
1000
1200
1400
Signal (ionization electrons)
20000
17500
15000
12500
10000
7500
5000
2500
0


gauss sim:2D/SP:2D, event 0, APA 0, U plane
Relative channel number
0
20
40
60
80
100
120
140
Relative ticks number
0
25
50
75
100
125
150
175
Signal (ionization electrons)
20000
17500
15000
12500
10000
7500
5000
2500
0


splat splat, event 0, APA 0, U plane
Relative channel number
0
20
40
60
80
100
120
140
Relative ticks number
0
25
50
75
100
125
150
175
Signal (ionization electrons)
20000
17500
15000
12500
10000
7500
5000
2500
0

Summary of signals from ADC

1. 2D Fourier transform of ADC: $\mathcal{M} \equiv \mathcal{F}(M)$
2. Deconvolve to get an average pre-signal: $\tilde{S}' = \mathcal{F}^{-1}(\mathcal{M}/\mathcal{R})$
3. Final signal-ROI selection in time-domain: $\tilde{S} = roi(\tilde{S}')$

Bonus from the signal processing

During decon, a low-pass filter is applied (250 kHz cut off)

- By Nyquist/Shannon the data become 4× **oversampled**.
- We may reap benefit by applying 4× rebinning with **no info lost**.

In addition to signals being sparse from signal-ROI, this rebin makes the signal data size much smaller than for ADC.

The "splat" tier

Given approximate cancellation for sim+sigproc with matched R described in the aside, we may consider to produce sim+sigproc-like tiers with `DepoSplat`.

- Far faster and approximately the same result as full sim+sigproc, **but**,
- `DepoSplat` has **no noise model**! While signal-ROI on $\mathbb{R}$ does a good job to remove the amplified low-frequency noise, it is imperfect.
  - Residual amplified noise in signals show as "speckle" on induction planes. `DepoSplat` does not and can not really model this.
- `DepoSplat` **lacks discretization** due to ADC digitization
  - In reality, that `floor()` is right between `decon()` and `conv()`.
  - It manifests as "jagged" tails of the final signals.
  - Only known option currently is to add some ad-hoc fakery.

Let's come back to this tier later as possible way to use LS4GAN for reconstruction.

Select LS4GAN Domain

So, the best input domain for LS4GAN is that of estimated average signal,

$$\tilde{S}!$$

Production processing

The `toyzero` package has been upgraded to produce $\tilde{S}$ for:

real signal from 2D simulation + q1D sigproc

fake signal from q1D simulation + q1D sigproc

Optionally it may also produce:

ADC save intermediate q1D and 2D ADC level

splat run `DepoSplat` as a side process

Production data sizes

ProtoDUNE-SP (6 APAs) in MBs/(four flavors of event)

[MB]	ADC	Signal	Splat
frame	198	11	5.4
images	111	6.4	1.8

- The **ADC** tier datasets are in **two flavors**: q1D and 2D.
- The **signal** tier datasets come in **four flavors** from outer-product of R used for (sim,sigproc) : (q1D,2D) $\otimes$ (q1D,2D).
- All are compressed (bzip2), "images" are smaller than frames due to $4\times$ rebin.
  - Hmm, actually this is a bug for the ADC "images" datasets!
- We may decide to not save: ADC (all), Splat (all), frames (all) and (q1D,2D) and (2D,2D) sim+sigprocs Signals.
  - or, we may save "frames" (6 files per **job**) and do the "split" into individual "image" files offline (6×3 files per **event**). Split job is $\approx$ 1s per APA-event.

Some possible production numbers

Based on a wild guess for how many events needed:

- 10k events per sim+sigproc "flavor".
- $\times 2$ flavors sim+sigproc: (q1D+q1D) and (2D+q1D).
- 300 CPU-seconds sim+sigproc per event.
- 32 GB of signal images or 55 GB of signal frames.
- 1600 CPU-hours
- Perhaps, choose 1 hour per job.