smartpixels

Data Reduction in Sensor Frontends for Improved Physics Capabilities at Colliders

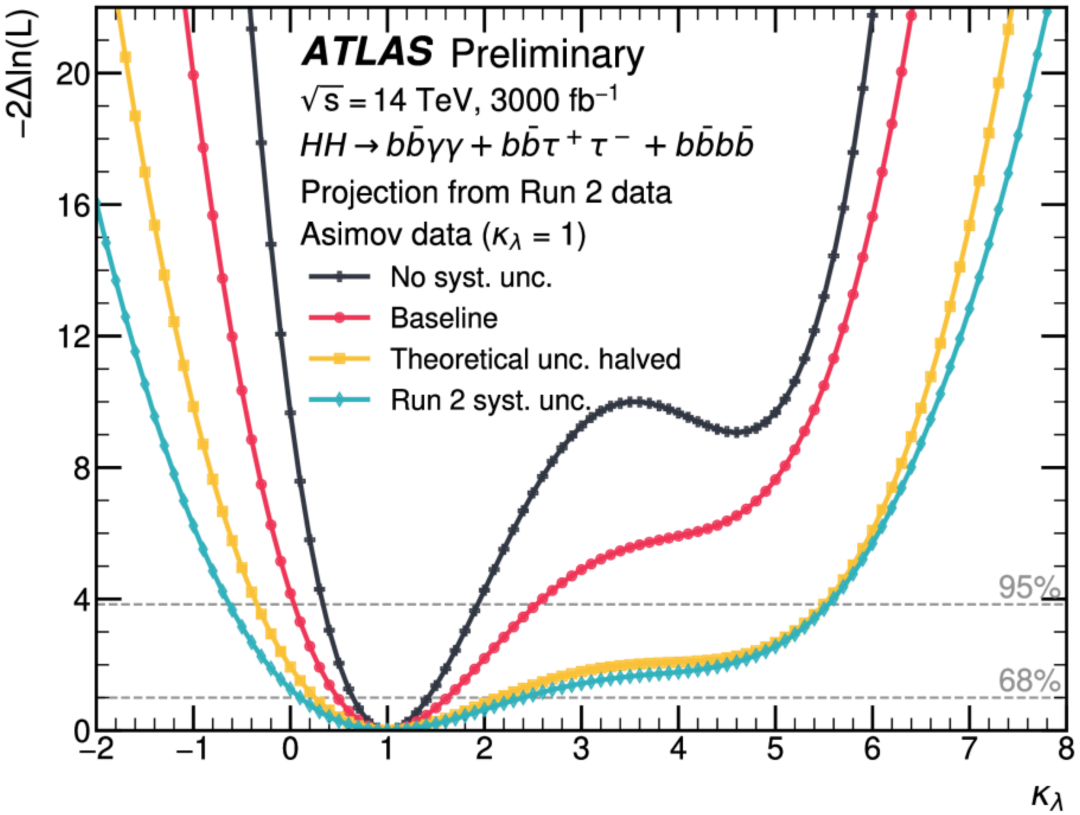
Lindsey Gray, on behalf of the smartpixels collaboration

ML4FE Workshop 2025

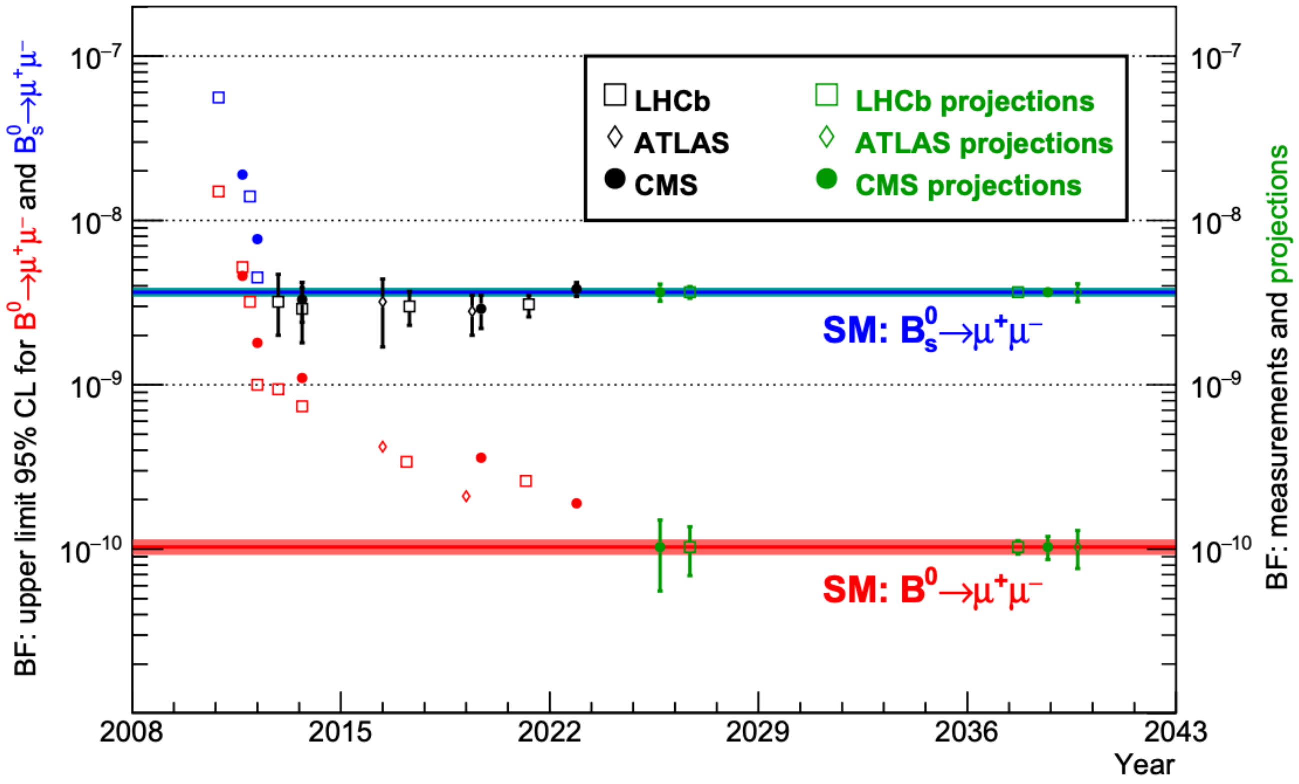
20 May 2025

Two Major Physics Opportunities of the HL-LHC

ATL-PHYS-PROC-2024-098

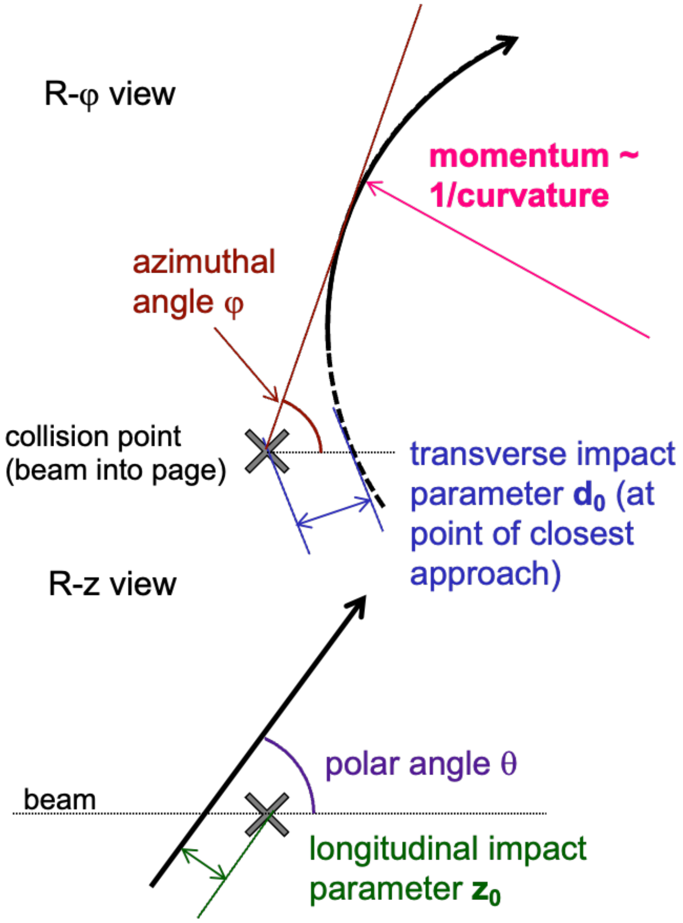


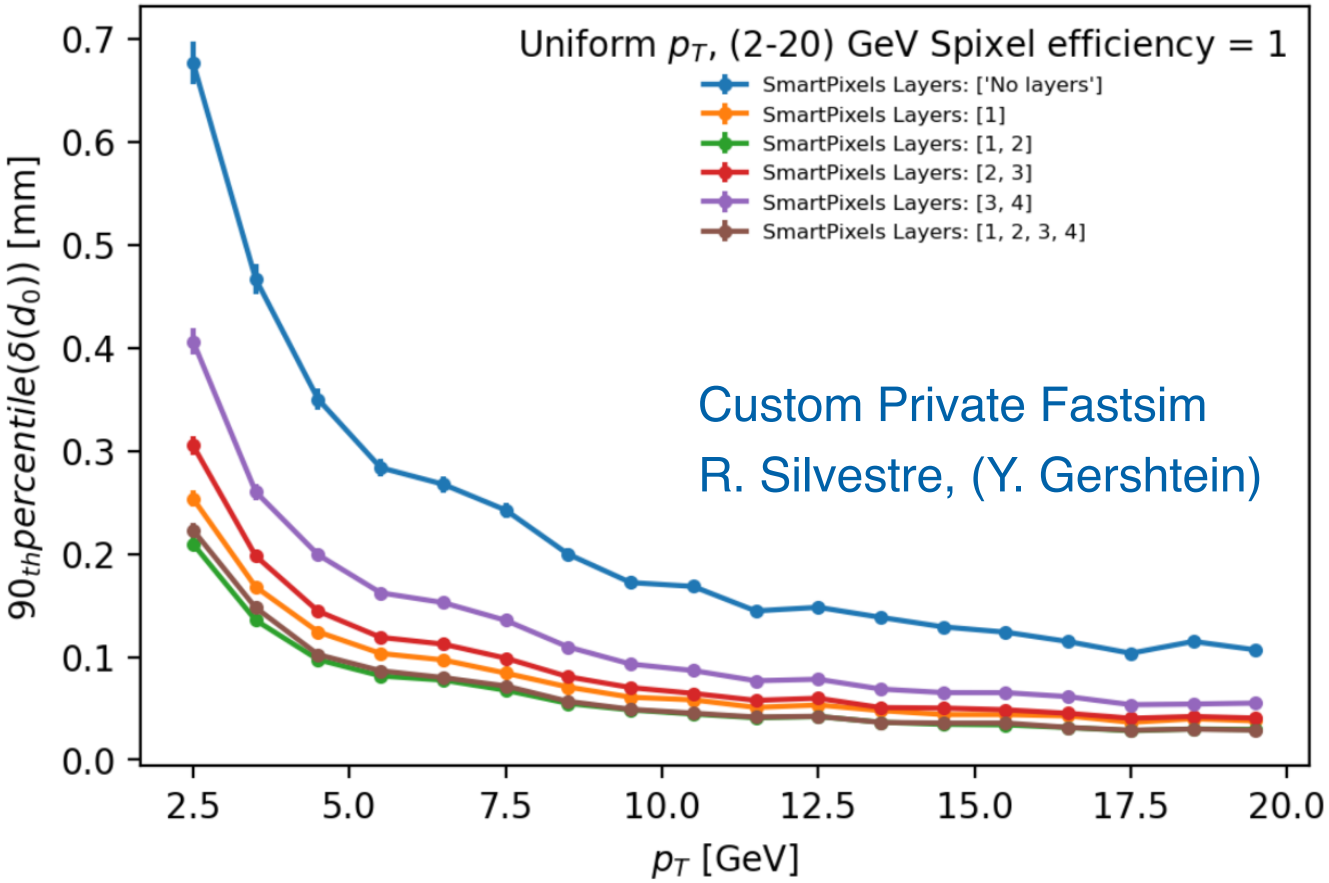
https://arxiv.org/pdf/2405.07693



- $B^0_{(s)}$ to muons measurements of HL-LHC are powerful tests of SM
  - B^0 lifetime is 1.445 +/- 0.018 ps, mass 5.3 GeV, large displacements, low pT tracks
  - Limited at HL-LHC by trigger thresholds and low pT d0 resolution in trigger
- The observation of the Higgs self-coupling at HL-LHC will stand as a guiding measurement for HEP until the pCOM 10 TeV Machines
  - Higgs Factories may provide some gains but…
  - No precision measurement of λ until FCChh / MuC - the better part of a century
  - We should do the best that we can with HL-LHC data - it is manifestly worth it

An Aside on Transverse Impact Parameters

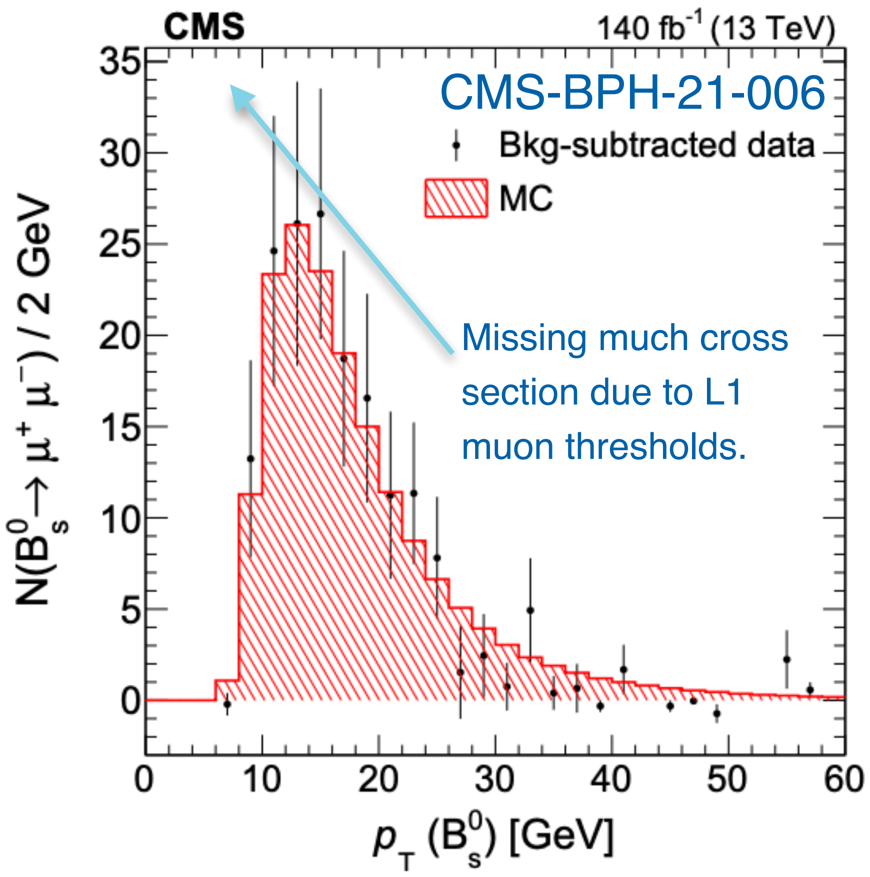

R-φ view
momentum ~ 1/curvature
azimuthal angle φ
collision point (beam into page)
transverse impact parameter $\mathbf{d_0}$ (at point of closest approach)
R-z view
polar angle θ
beam
longitudinal impact parameter $\mathbf{z_0}$



Custom Private Fastsim
R. Silvestre, (Y. Gershtein)

- Transverse impact parameter resolution heavily influenced by pixel cluster precision and multiple scattering
  - LHC detector trackers are *heavy*
- Without innermost measurements:
  - extrapolated track fit has enormous position errors at low pT
  - Physics improvements only at large lifetimes / high transverse boost
  - *Poor precision* for physically interesting phase space

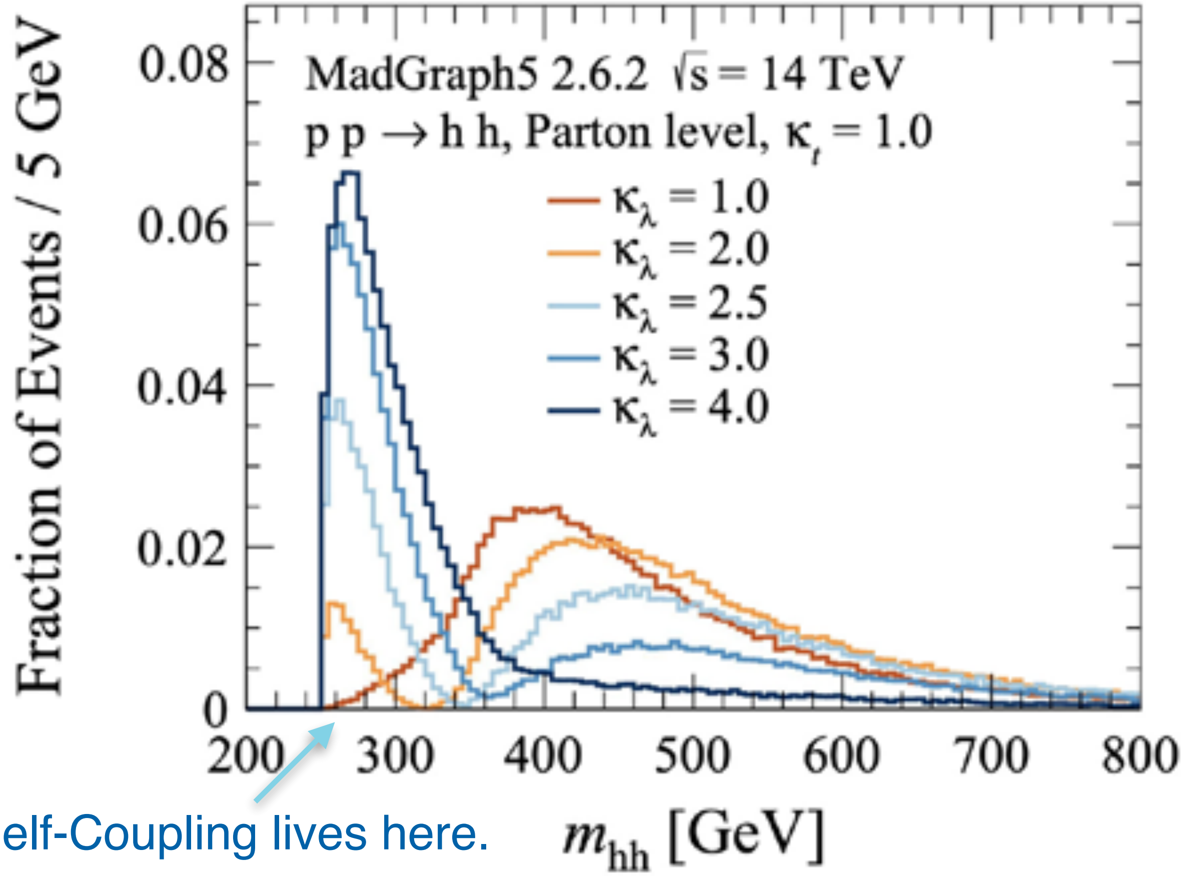
The Physics Limitations of Using “Only” the Outer Tracker



CMS-BPH-21-006

Missing much cross section due to L1 muon thresholds.

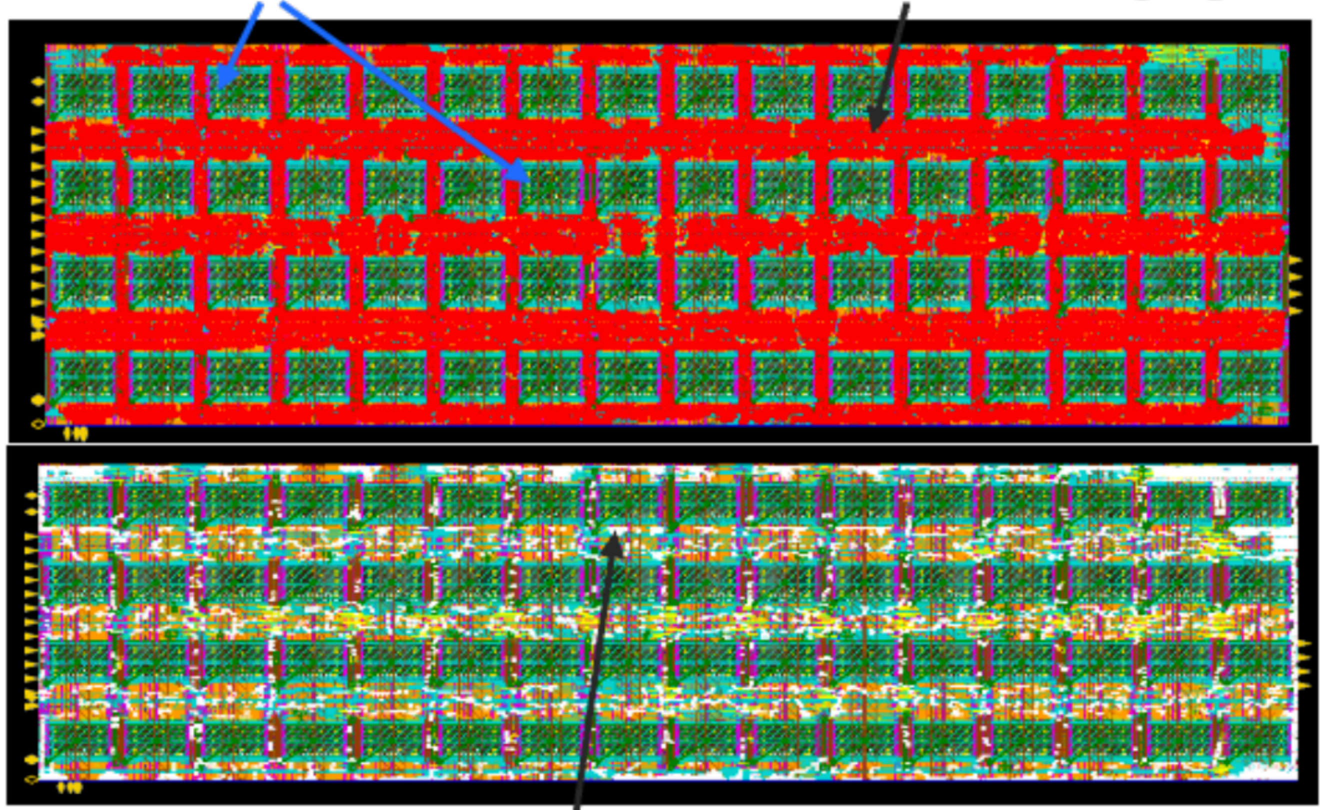
https://doi.org/10.1142/S0217751X24430036



The Higgs Self-Coupling lives here.

- Certainly, there is no “only” here
  - OTTF ingests 40 tbps of information and is a triumph of systems design
- However, the system’s location can limit its physics performance:
  - We miss the *most physics-rich* kinematic regions at the HL-LHC due to trigger rate
  - Even with the recent advances in L1 b-tagging, there is substantial room for gain
  - If we have online high-resolution transverse impact parameter determination we enable real time b-physics analysis and offline-like heavy flavor tagging for jets
    - Remove/improve baseline trigger acceptance for these processes (the *first* cut of any analysis)

How to get there? Maybe smartpixels!



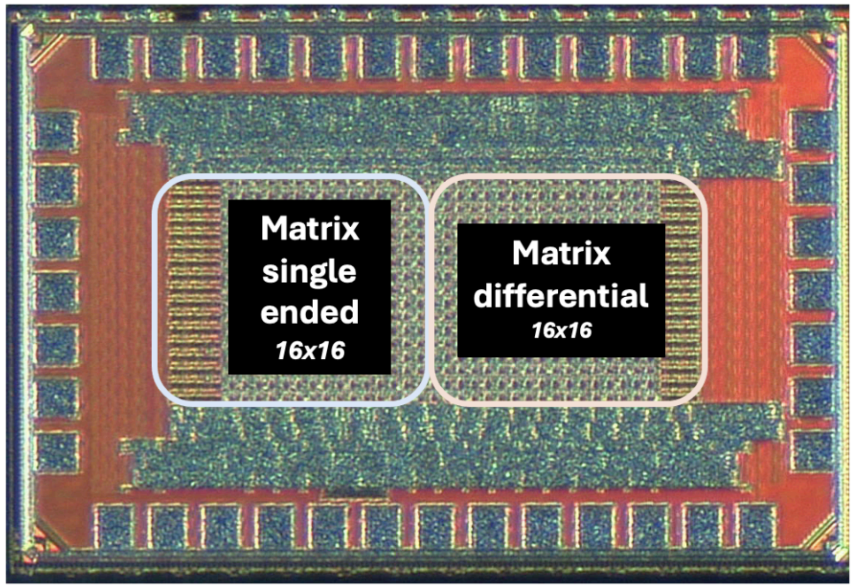
Signal processing clusters

Digital Momentum classifier distributed in-between sensing regions

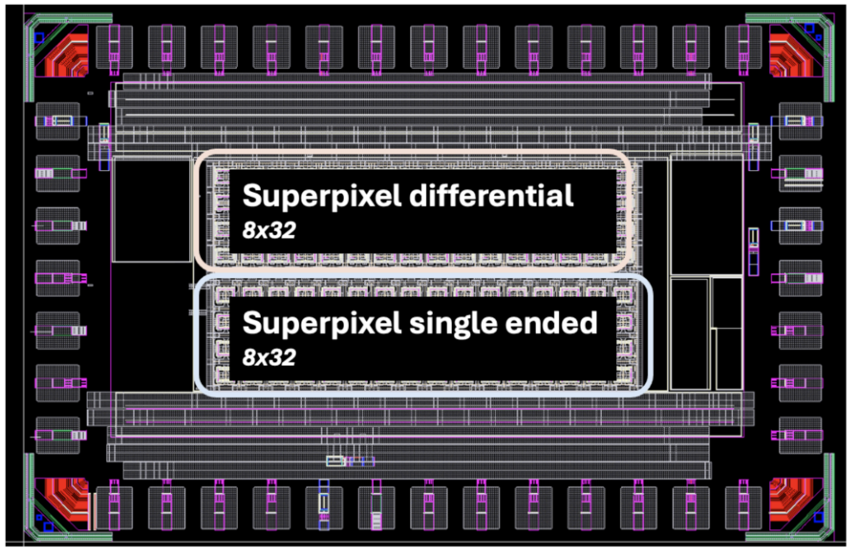
Reprogrammable weights distributed across the matrix (highlighted in white)

See Earlier Talk by B. Parpillon

a. First ROIC chip prototype



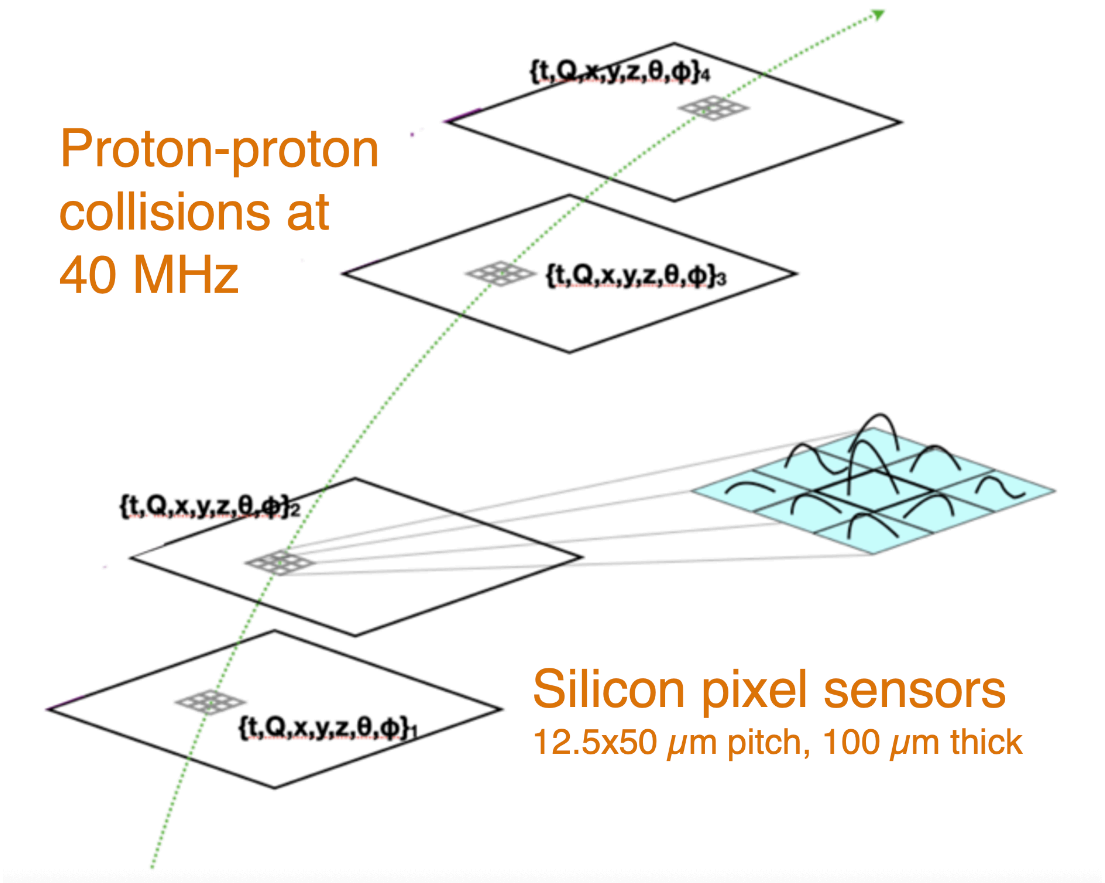
b. Second ROIC chip prototype



- We are presently testing prototypes of a configurable smartpixels filter chip
- We are working with the VIAS project to implement an in-ASIC regression
- Submitted, waiting to hear back on Hardware Aware AI FOA from last year
  - Would allow us to produce a bump-bondable prototype filter + regression chip
  - Crossing our fingers with everyone else here who submitted :-)

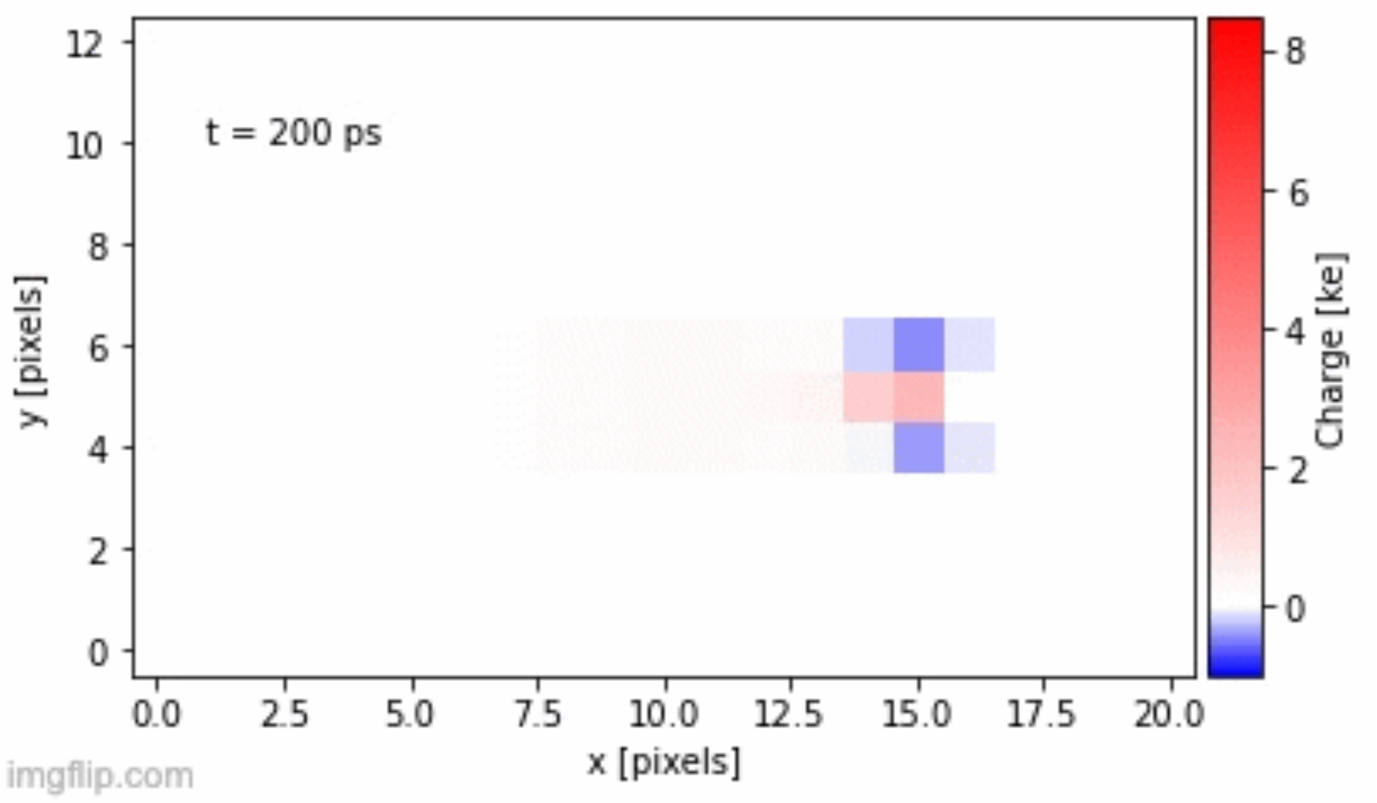
Smartpixels Sensor Dataset

We made it public, please use it!
https://zenodo.org/records/7331128



Proton-proton collisions at 40 MHz

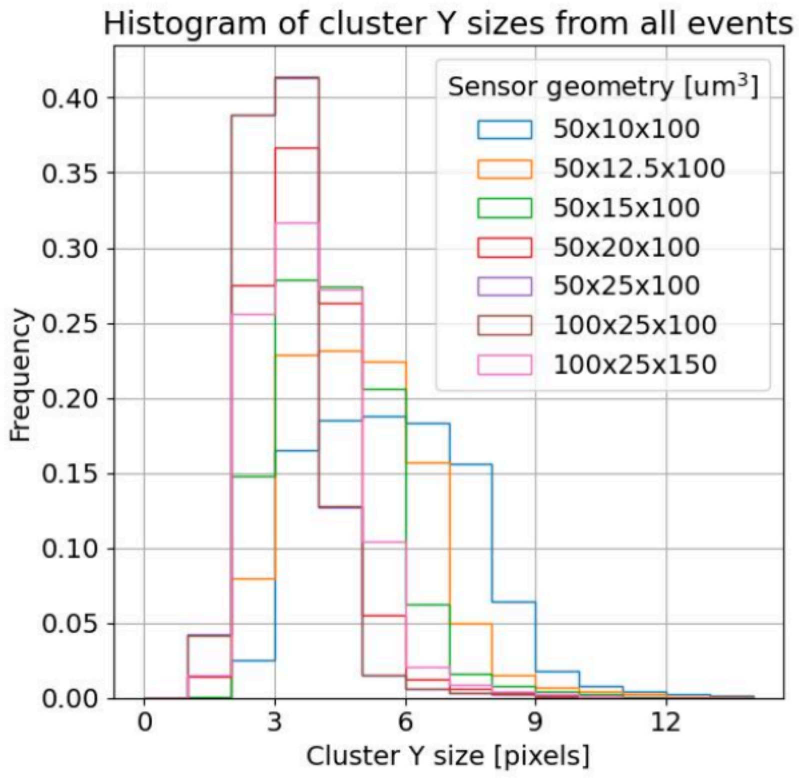
Silicon pixel sensors
12.5x50 μm pitch, 100 μm thick



- For the studies so far we've used a 21x13 pixel array with 20 samples with 200 ps spacing, pixel size is 50um x 12.5um
  - PixelAV is used to perform sensor simulation, momenta are taken from real data
  - The "readout" is a running integral of the pulse (~idealized CSA)
- For realism we take fewer time slices and digitize the inputs
  - For the filter results to follow we use 2-bit digitization
  - For the regression results we use a 4-bit digitization, we are working on 2-bits

Sensor / ML Performance Optimization (Filter)

Danush Shekar



Cluster y-size for all sensors

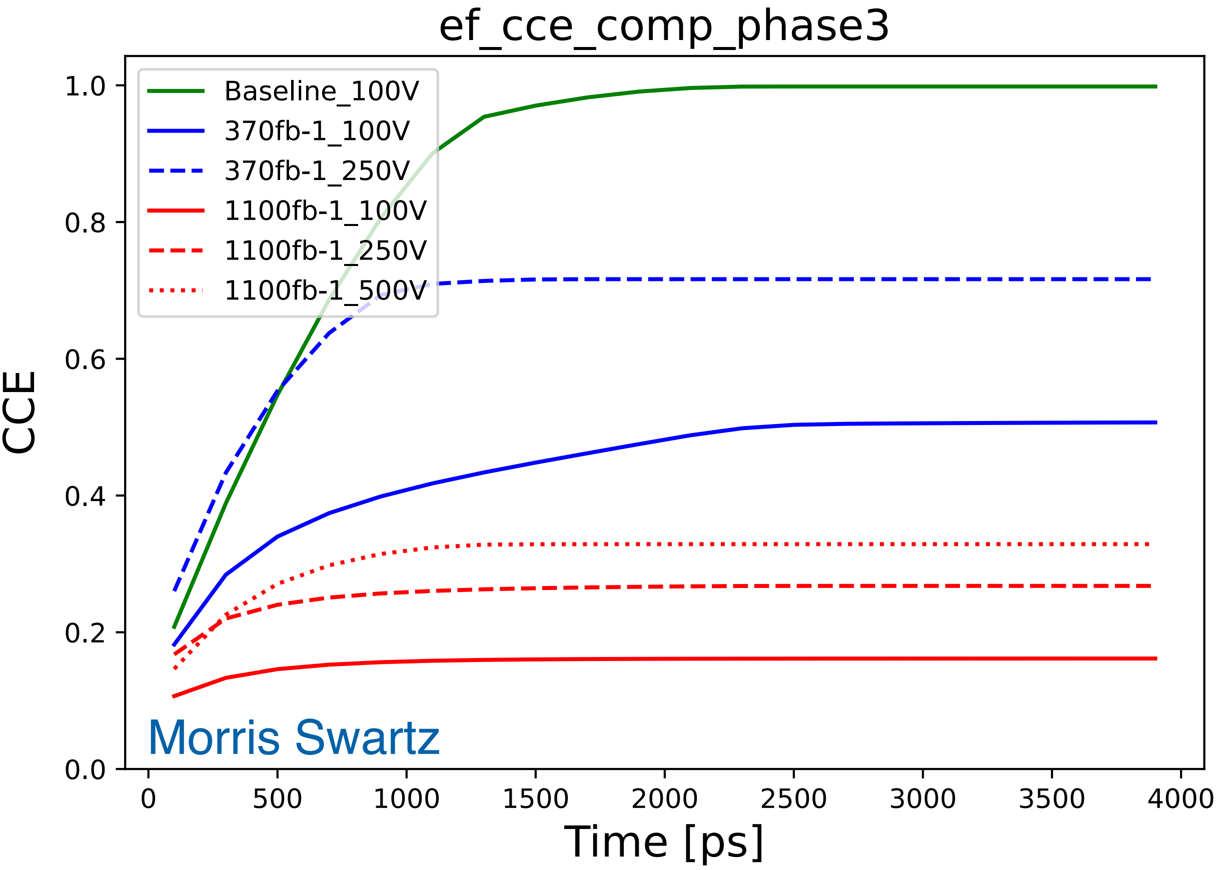
Nominal Phase 2 Pixel

Sensor geometry [um^3]	Bias voltage [V]	Signal efficiency	Data reduction
50 X 10 X 100	100	95.4 ± 0.5	33.1 ± 1.0
50 X 12.5 X 100	100	93.9 ± 0.5	33.1 ± 0.9
50 X 15 X 100	100	93.3 ± 0.5	30.7 ± 0.9
50 X 20 X 100	100	91.2 ± 0.9	28.4 ± 0.9
50 X 25 X 100	100	88.3 ± 0.7	27.3 ± 0.8
100 X 25 X 100	100	88.6 ± 0.9	26.9 ± 1.0
100 X 25 X 150	175	91.9 ± 0.7	29.7 ± 1.0

- Recently we have been focusing on making sure we're optimizing the sensor in the context of ML performance
  - Interesting to note that as we reduce pixel size we are able to discriminate low pT clusters better
  - There is still more information the finer we spatially sample the charge deposition!
  - Perhaps just as interesting, it still works ~well with Phase 2 sensor specs!
- Must balance pitch with data rate of reading out all these pixels and matters of assembly
  - The codesign of an integrated sensor system goes much deeper than the neural network!

First Looks at Radiation Damage Effects (Filter)

Danush Shekar



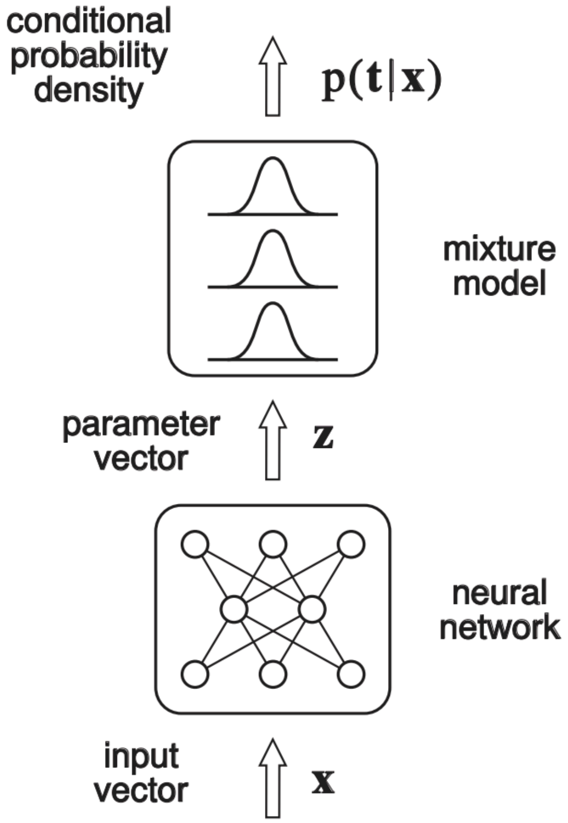
Charge Collection Efficiency

Morris Swartz

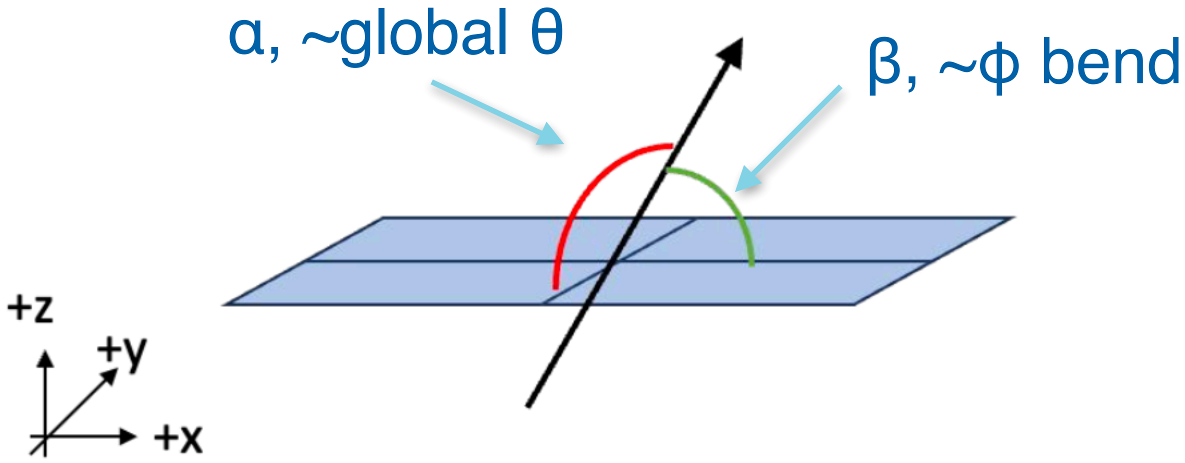
Irradiation level [fb^{-1}]	Bias voltage [V]	Signal efficiency	Data reduction
0	100	93.9 ± 0.5	33.1 ± 1.0
370	250	90.9 ± 0.7	29.7 ± 0.9
1100	500	87.1 ± 0.9	29.3 ± 1.0

- Similarly we have looked at the performance as a function of radiation exposure
  - While there is degradation in signal efficiency and background rejection it is not dire
  - Note: in order to ~maintain performance model retraining *is required*
- The more rapid charge collection from higher voltage operation reduces effect of Lorentz drift and decreases available information
  - The reduction in information is reflected in the model performance quite clearly

Regressing Track Parameters from Pixel Clusters



Bishop, 1994



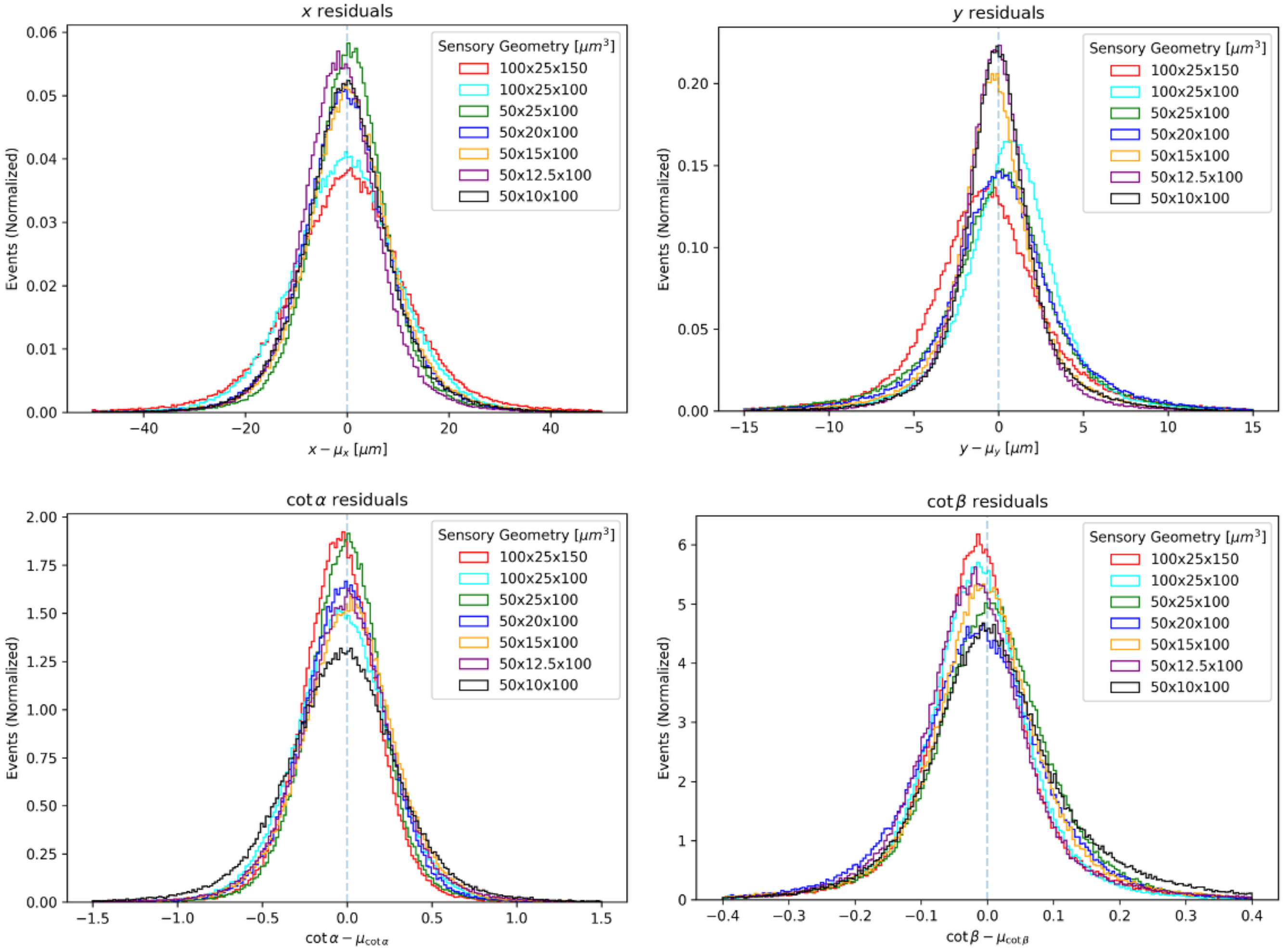
Network average estimated errors:

δα ~ 5 degrees, δβ ~ 2 degrees

- In synthesis: real time per-cluster estimates of track position and angle, including errors
  - We can do (very poor) tracking with this network
  - Initial networks are able to achieve II of 1 clock, latency of 2 clocks using a 2D CNN with separable convolutions to reduce total ops
  - Initial attempt took too much floorspace on chip, getting closer iterating on network architecture
- Assuming each track has some resolution we can describe each track as a sample from a 4D gaussian with parameter means and covariances
  - Train by minimizing likelihood, using NN to predict parameters of gaussian
  - Using QKeras -> HLS4ML -> Siemens Catapult for codesign pipeline (also Vivado for very early tests)

Regression Performance

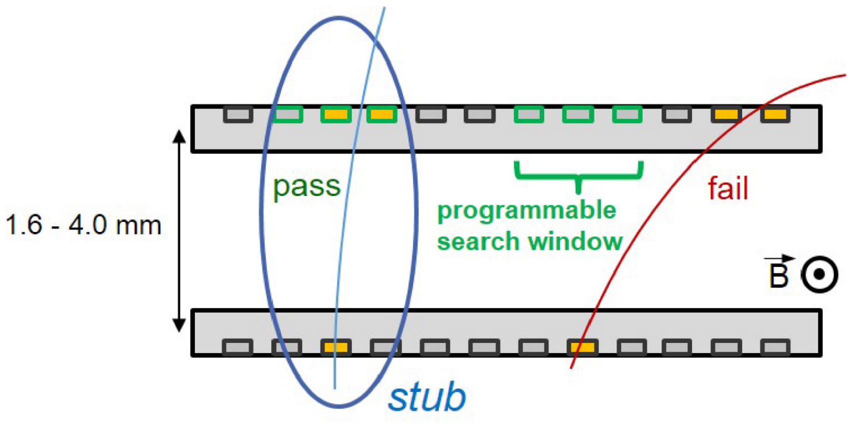
David Jiang

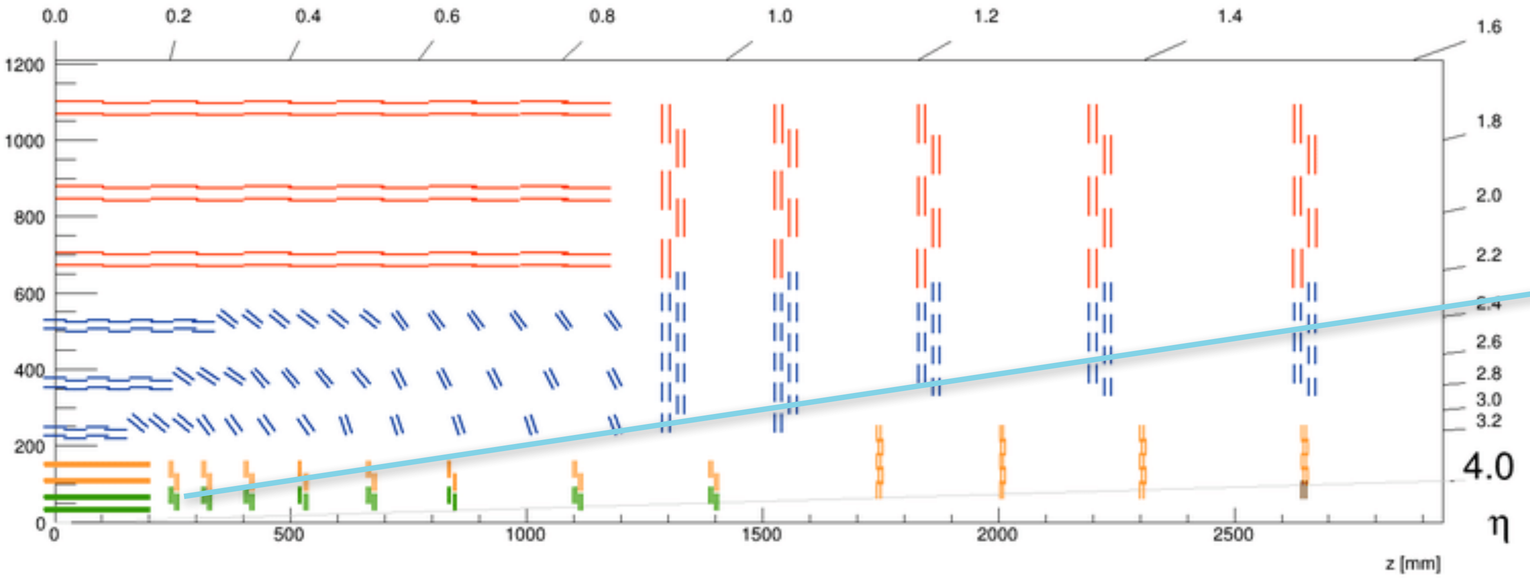


	X pull fit		Y pull fit		Cot(α) pull fit		Cot(β) pull fit	
Sensor geometry [um^3]	μ	σ	μ	σ	μ	σ	μ	σ
50 X 10 X 100	0.06	0.94	-0.19	0.86	0.16	0.72	-0.03	0.71
50 X 12.5 X 100	0.03	0.97	-0.09	0.95	-0.18	0.75	0.17	0.94
50 X 15 X 100	-0.08	0.85	0.07	0.87	0.02	0.80	0.04	0.93
50 X 20 X 100	0.03	0.90	-0.16	0.87	0.02	0.87	-0.17	0.84
50 X 25 X 100	0.04	0.88	0.10	0.89	0.02	0.82	0.05	0.75
100 X 25 X 100	0.07	0.98	-0.23	0.86	0.02	0.83	0.08	0.81
100 X 25 X 150	0.25	0.80	0.01	0.72	-0.05	0.84	-0.05	0.68

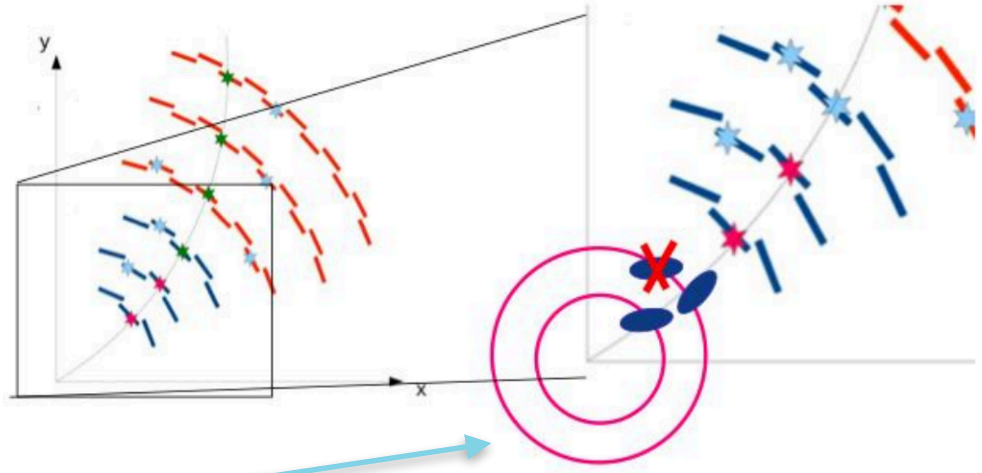
- We also studied the performance of the regression network across a variety of sensor pitches
  - We find that we are able to maintain decent position and angle reconstruction across all pixel pitches
  - Pulls of error distributions are reasonable but indicate over estimated errors

Using smartpixels data in a detector system





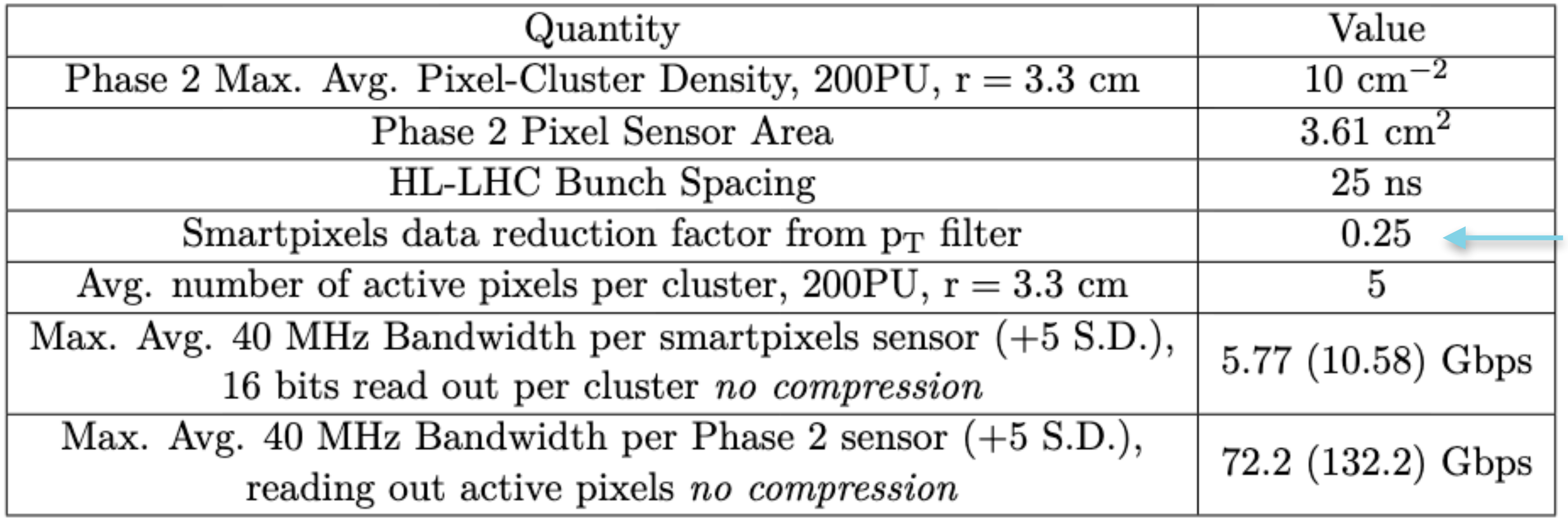
Start from OT tracks



Use pixel directional information to improve hit-match quality. Reconstructed d0 resolution easily degraded by incorrect matches.

- We get roughly 2-3 degree resolution in the azimuthal direction
  - “Inside-out” tracking will not work very well, combinatorics too high
  - However we can use the outer tracker tracks to identify regions of interest and then the angular reconstruction from smartpixels to match the extrapolated track
- With more processing on detector we could try to find “pixel seeds” instead
  - More clean and pure, but this would require on detector track finding to deal with combinatorics

Estimates of Data System Throughput

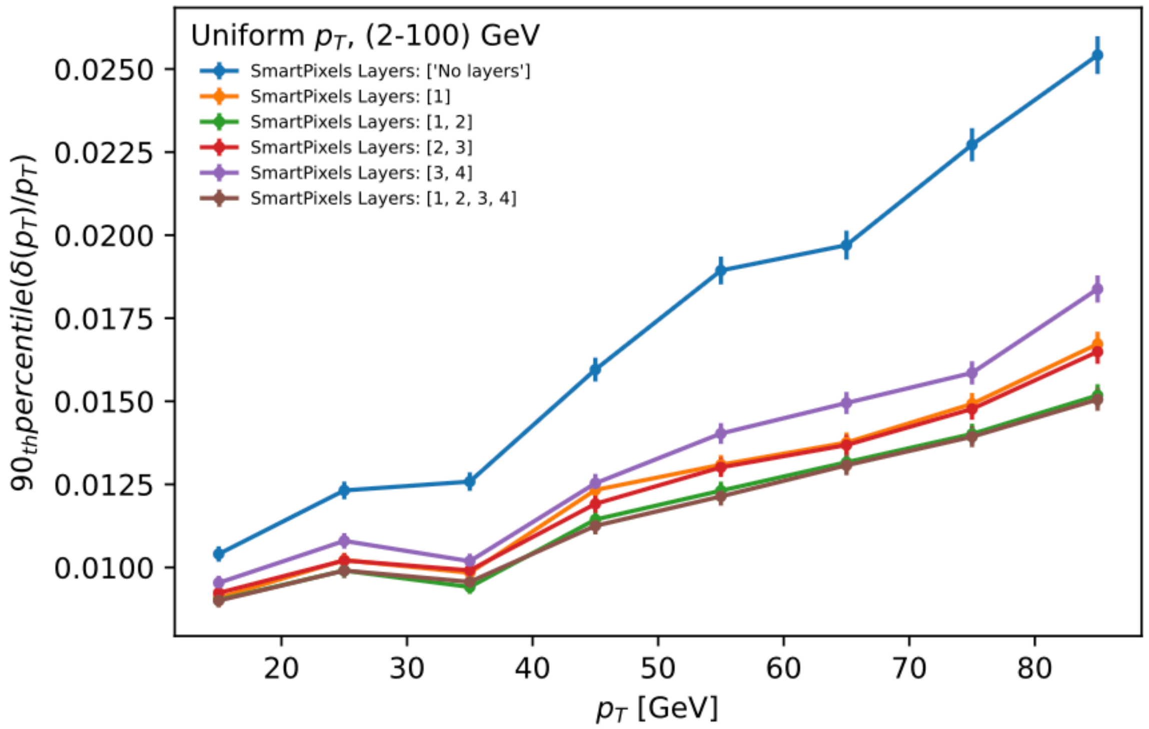


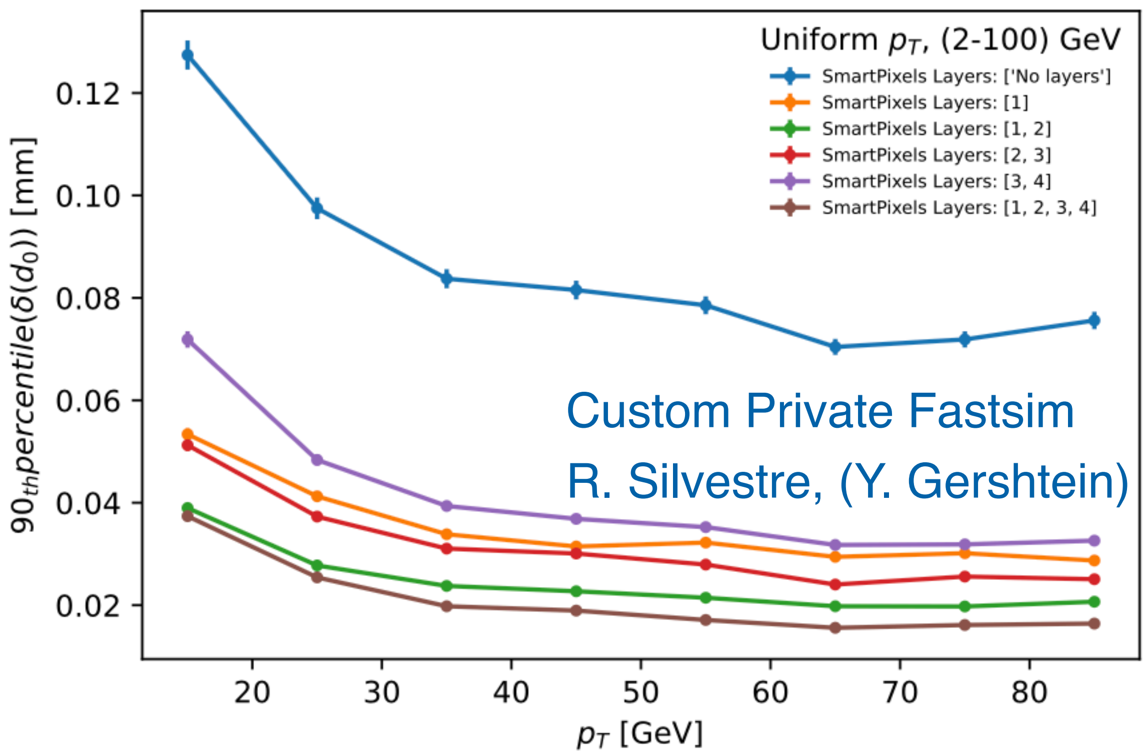
Quantity	Value
Phase 2 Max. Avg. Pixel-Cluster Density, 200PU, $r = 3.3$ cm	10 cm^{-2}
Phase 2 Pixel Sensor Area	3.61 cm^2
HL-LHC Bunch Spacing	25 ns
Smartpixels data reduction factor from p_T filter	0.25
Avg. number of active pixels per cluster, 200PU, $r = 3.3$ cm	5
Max. Avg. 40 MHz Bandwidth per smartpixels sensor (+5 S.D.), 16 bits read out per cluster *no compression*	5.77 (10.58) Gbps
Max. Avg. 40 MHz Bandwidth per Phase 2 sensor (+5 S.D.), reading out active pixels *no compression*	72.2 (132.2) Gbps

NB: this assumes improved performance from a future model

- This calculation makes it abundantly clear why streaming pixels is difficult
  - Without ML in FE readout requirements are beyond the next decade+ of rad-hard telecommunications products in terms of bandwidth, power usage, and waste heat
  - Using a ML based compression a two-layer smartpixels system throughput is ~7 Tbps
    - Equivalent to ~another layer of the outer tracker (note: this still *requires* rad-hard silicon photonics)
  - Note: This assumes we do not read out errors and only (x, y, phi) of clusters in 16 bits
- This will likely require an additional rack of processing equipment
  - Standard pixel readout + 40 MHz L1 readout + interface to extrapolate L1 OTTF tracks and then align and refit with pixel data
  - Using Phase 2 upgrade technologies O(1000) fibers, O(20) new boards, and entirely new pixel DAQ
    - Less assuming technological improvements since then, e.g. silicon photonics, latest FPGAs
  - So - the chip is not the whole story but it is the critical enabling technology!

Suggestions of Physics Performance (I)



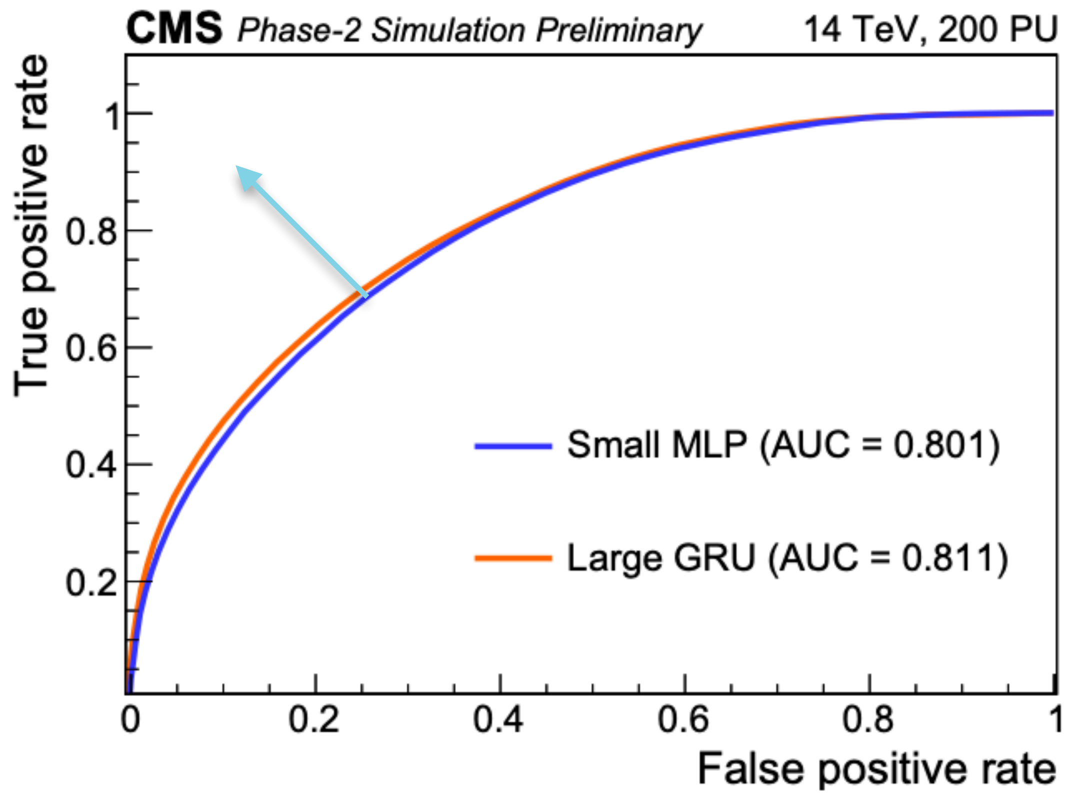


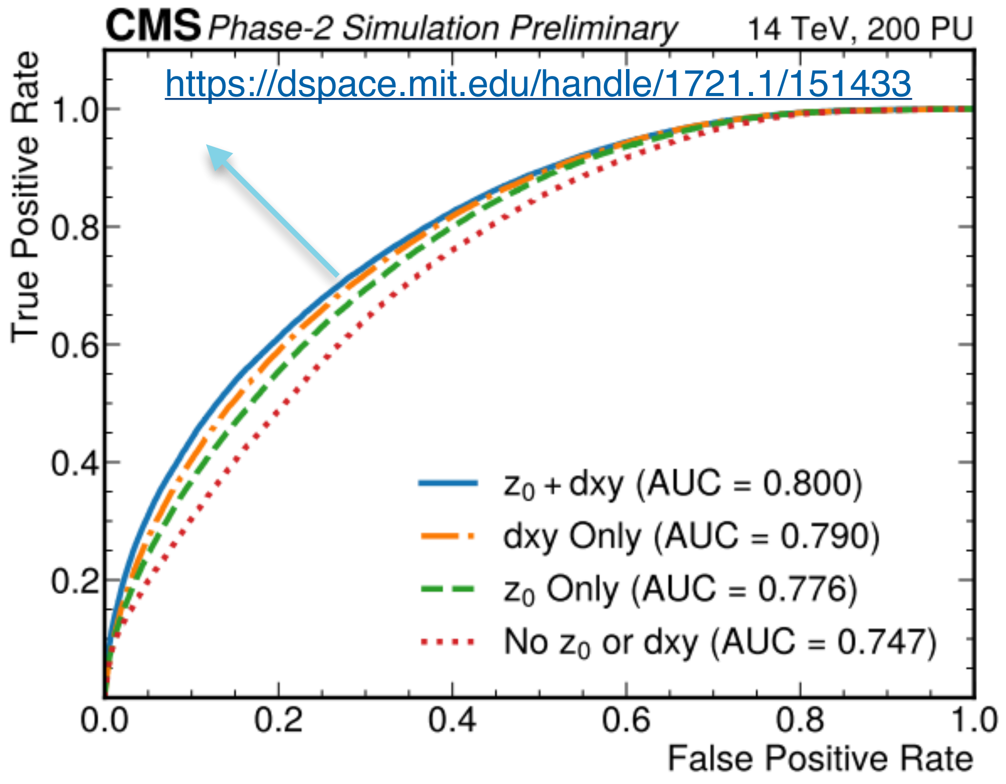
Custom Private Fastsim

R. Silvestre, (Y. Gershtein)

- Fast simulation results give us an indication of physics performance and detector optimization
  - Assumes 90% layer efficiency, expected CMS Phase 2 hit reconstruction performance
- Interesting considerations for detector placement
  - Very similar performance for layers [1,2] vs. [2,3] of pixel detector
  - Nearly 2x less required data throughput in 2nd layer
  - Radiation damage performance to consider as well, performance changes quickly

Suggestions of Physics Performance (II)

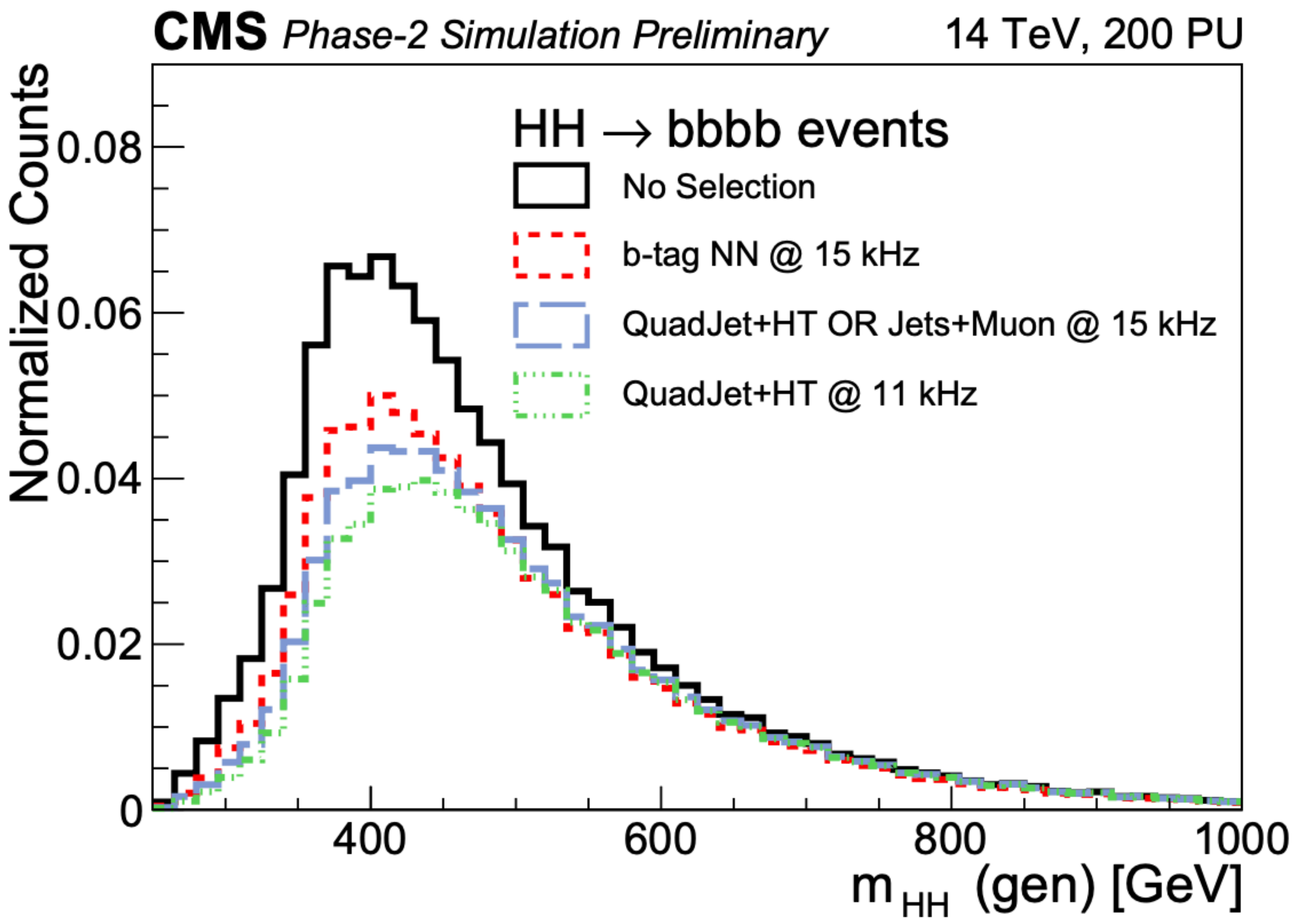


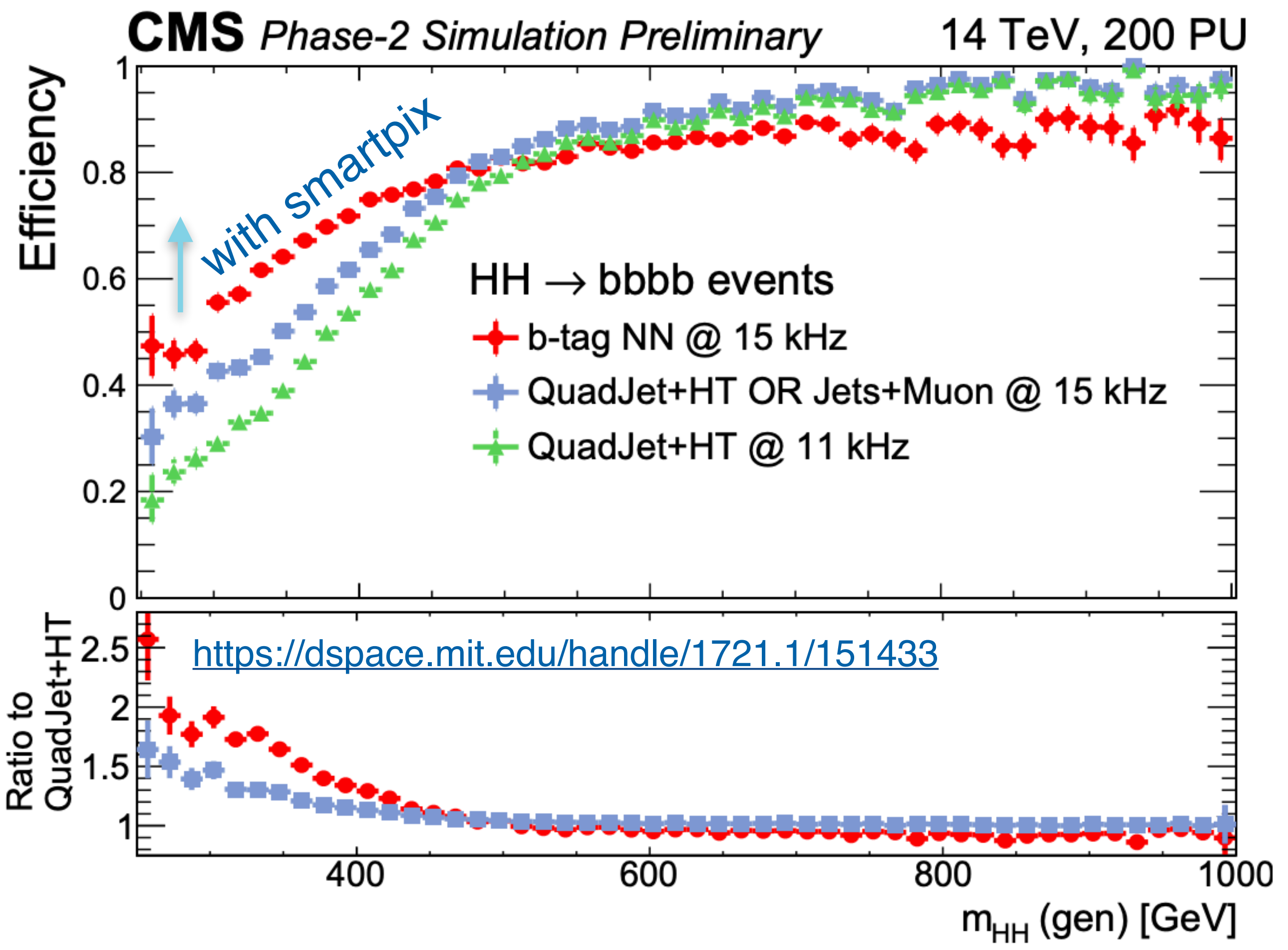


https://dspace.mit.edu/handle/1721.1/151433

- Pixel data stands to improve AUC to ~0.95 or better
  - Cannot show the actual plot in a workshop, it is CMS internal and preliminary
- However, a few interesting suggestions to make
  - d0, even from outer tracker alone, improves tagging performance
  - Improved d0 estimate including pixels will have even more effect
  - It will be interesting to understand how relative impacts change with improving quality

Suggestions of Physics Performance (III)





https://dspace.mit.edu/handle/1721.1/151433

- Recovering most of the remaining AUC will improve the kinematic acceptance in the most sensitive region to the Higgs Self-Coupling
  - Drastic signal efficiency increases at constant background rate
- Stand to recover majority of region between dotted red and black lines above
  - Reminder: above plots are outer tracker information *only*

Conclusions / Remarks

- This is all applicable to Future Higgs Factories and pCOM 10 TeV Machines
  - Discrimination of hits from beam-induced backgrounds compared to hard scatter
  - The MuC will need to use ML in a very different way because of the extreme occupancy
  - Sorry I didn't talk about this - let your imaginations guide you!

- Smartpixels, and ML in the frontend in general, is a powerful technology that if implemented may improve the scientific legacy of HL-LHC
  - If we can make it in time for Run 5 (the event horizon is quickly approaching)
  - Such a system could make available:
    - 40 MHz b-physics program with very little kinematic bias
    - Enhanced b-tagging in the level one track trigger with improved acceptance in the most physically relevant regions

- We hope to expand to other detector types and fully work out the use case in pixel systems at various collider types
  - Possibly conjoin with HGCAL ECON efforts, Dual-Readout Calorimetry, pixellated (tera-hit) calorimetry, multi-layer real time tracking

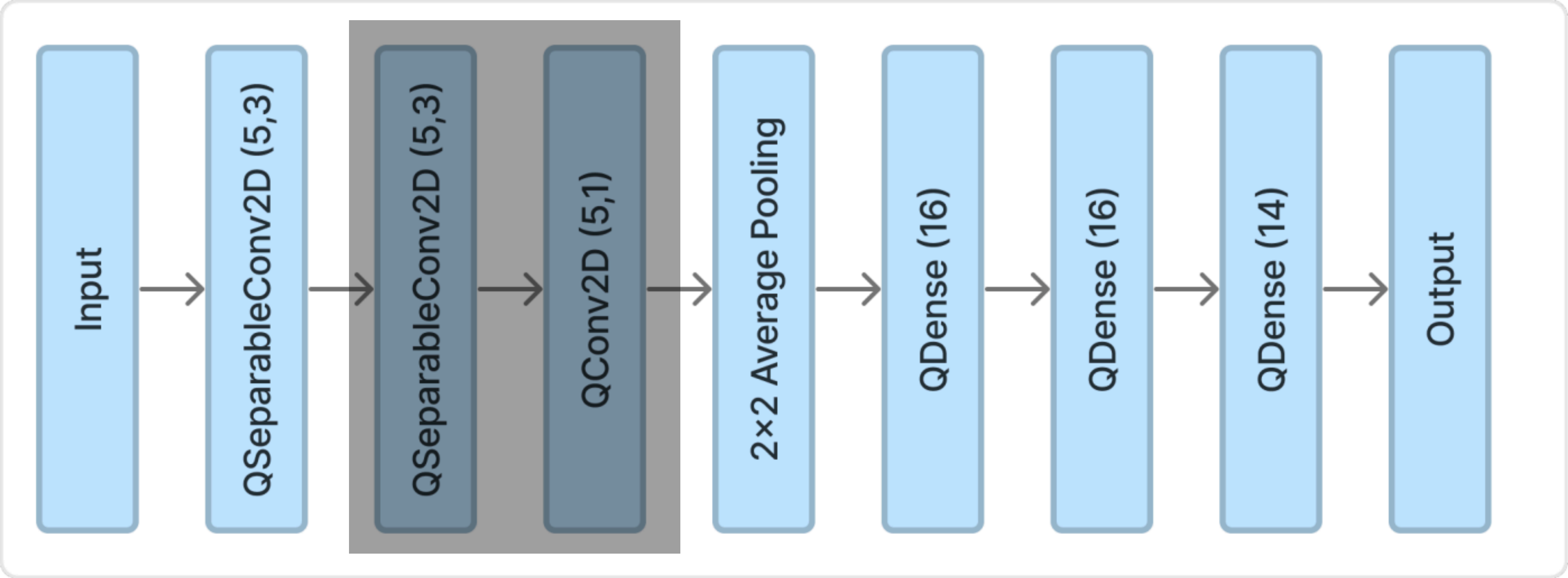
Extras

The Smartpixels Team

- Find us at: https://fastmachinelearning.org/smart-pixels/

- Team Members:
  - Daniel Abadjiev, Anthony Badea, Alice Bean, Douglas Berry, Arghya Ranjan Das, Jennet Dickinson, Karri DiPetrillo, Giuseppe Di Guglielmo, Farah Fahim, Abhijith Gandrakota, Lindsey Gray, Eliza Howard, James Hirschauer, David Jiang, Shruti R. Kulkarni, Carissa Kumar, Shiqi Kuang, Nicholas Manganelli, Miaoyuan Liu, Mira Littmann, Ron Lipton, Corrinne Mills, Petar Maksimovic, Aidan Nicholas, Mark S. Neubauer, Jannicke Pearkes, Benjamin Parpillon, Jieun Yoo, Nhan Tran, Ricardo Silvestre, Morris Swartz, Chinar Syal, Danush Shekar, Keith Ulmer, Dahai Wen, Ben Weiss, Mohammad Abrar Wadud, Aaron Young, Eric You, Albert Zhou

Regression Network Design



Input: 4 bit, eventually 2

Convolutions: 4 bit weights

Average, decoder, outputs: 8 bit weights

- Regression model is a relatively standard CNN architecture
  - We were able to aggressively quantize the convolutional layers
  - The decoder layers started to significantly degrade in performance below 8 bit weights
  - Since we *will* need to update network weights, and as of yet, we have not investigated the impact of radiation damaged sensor performance on networks pruned on fresh sensors, we do not consider pruning in the baseline
    - Rather we use it to inform us of the usefulness of layers and neuron multiplicity within layers of a non-pruned architecture
- We are investigating removing the grayed-out boxes as well as "projected" architectures to reduce the floor-plan size of the network further while retaining performance