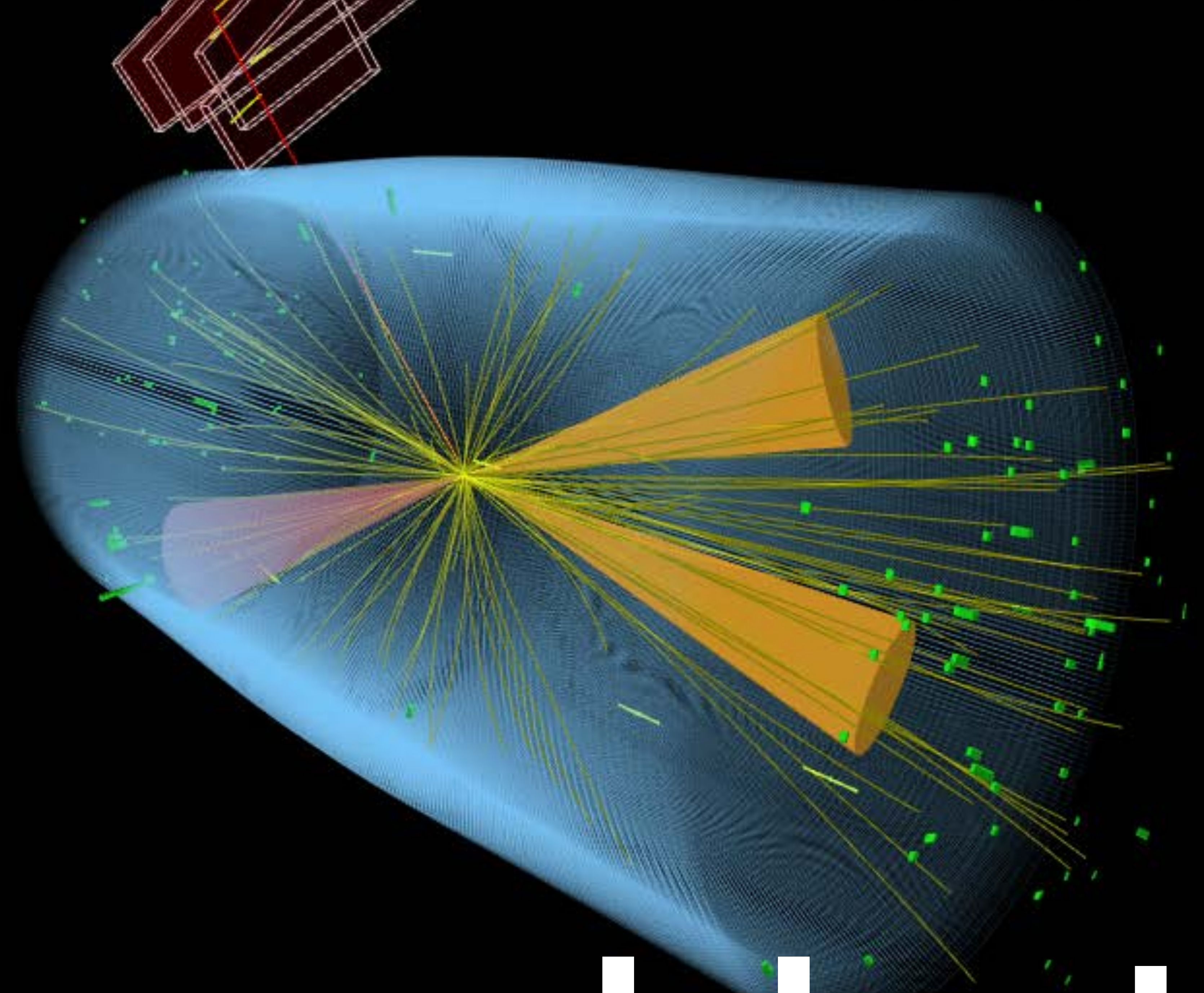


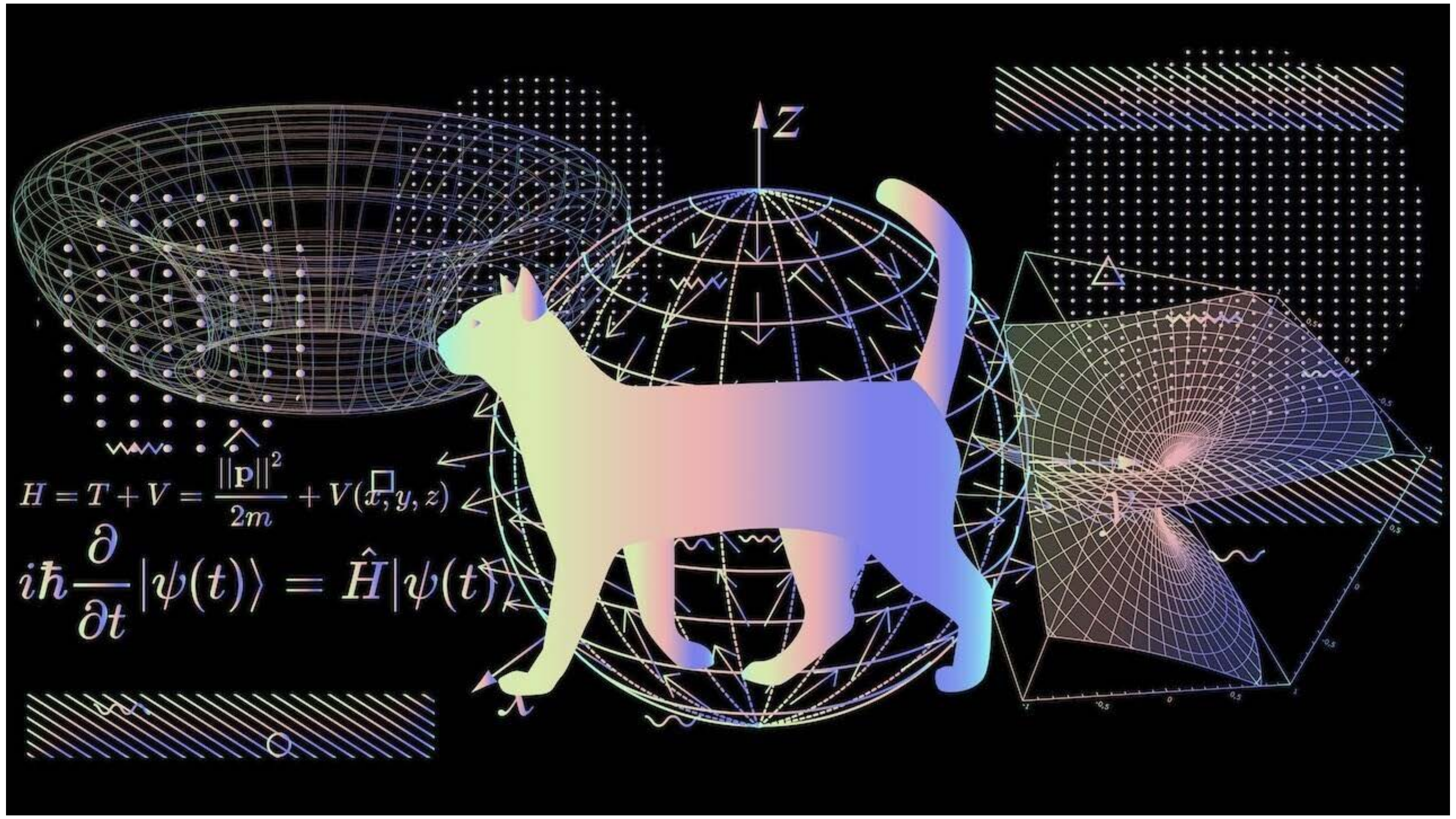
Foundation models at the edge for particle physics

IAIFI Symposium on Generative AI in the Physical Sciences

Thea Klaeboe Aarrestad (ETH Zurich)

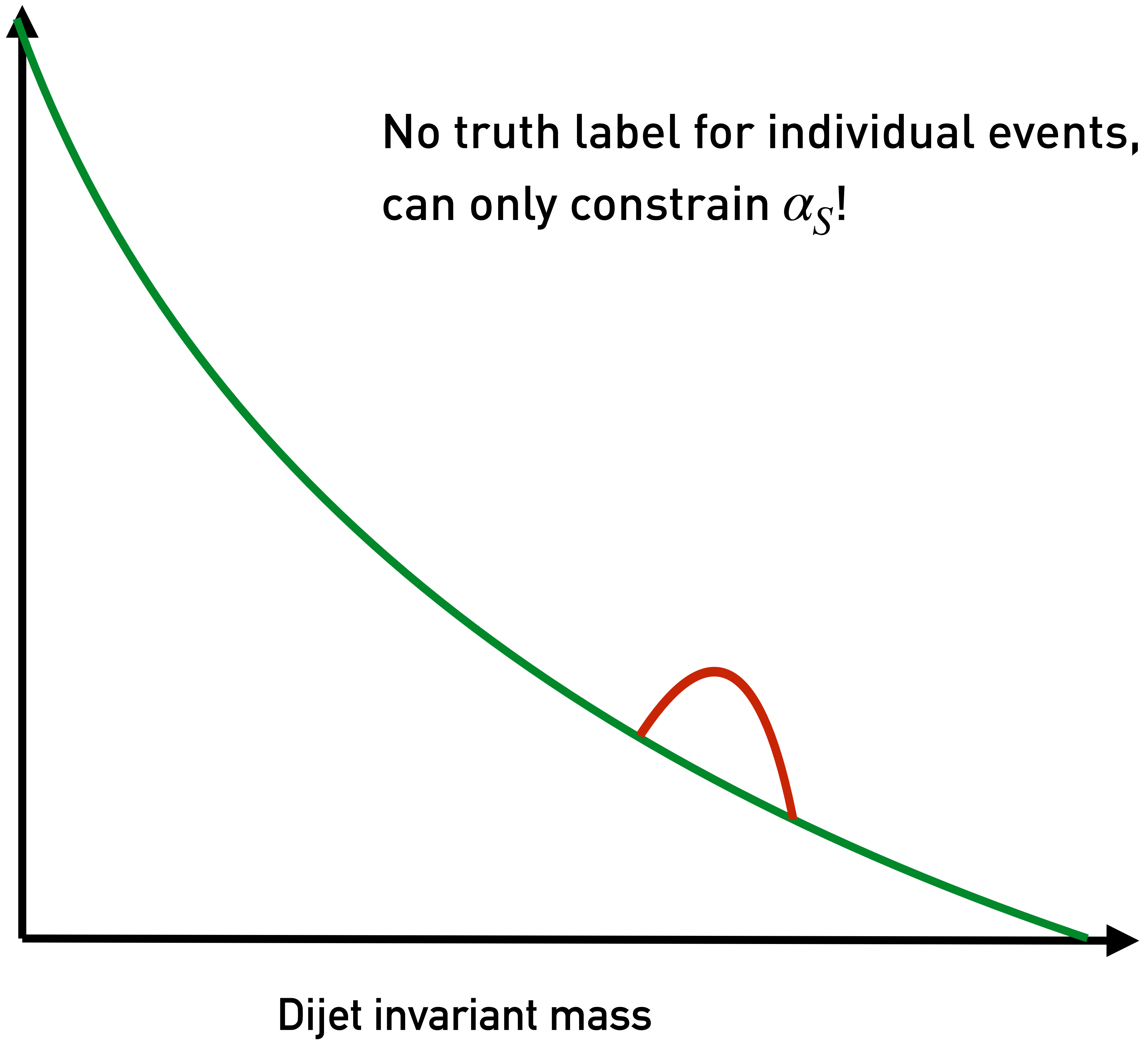
It's against physical law to annotate our data!

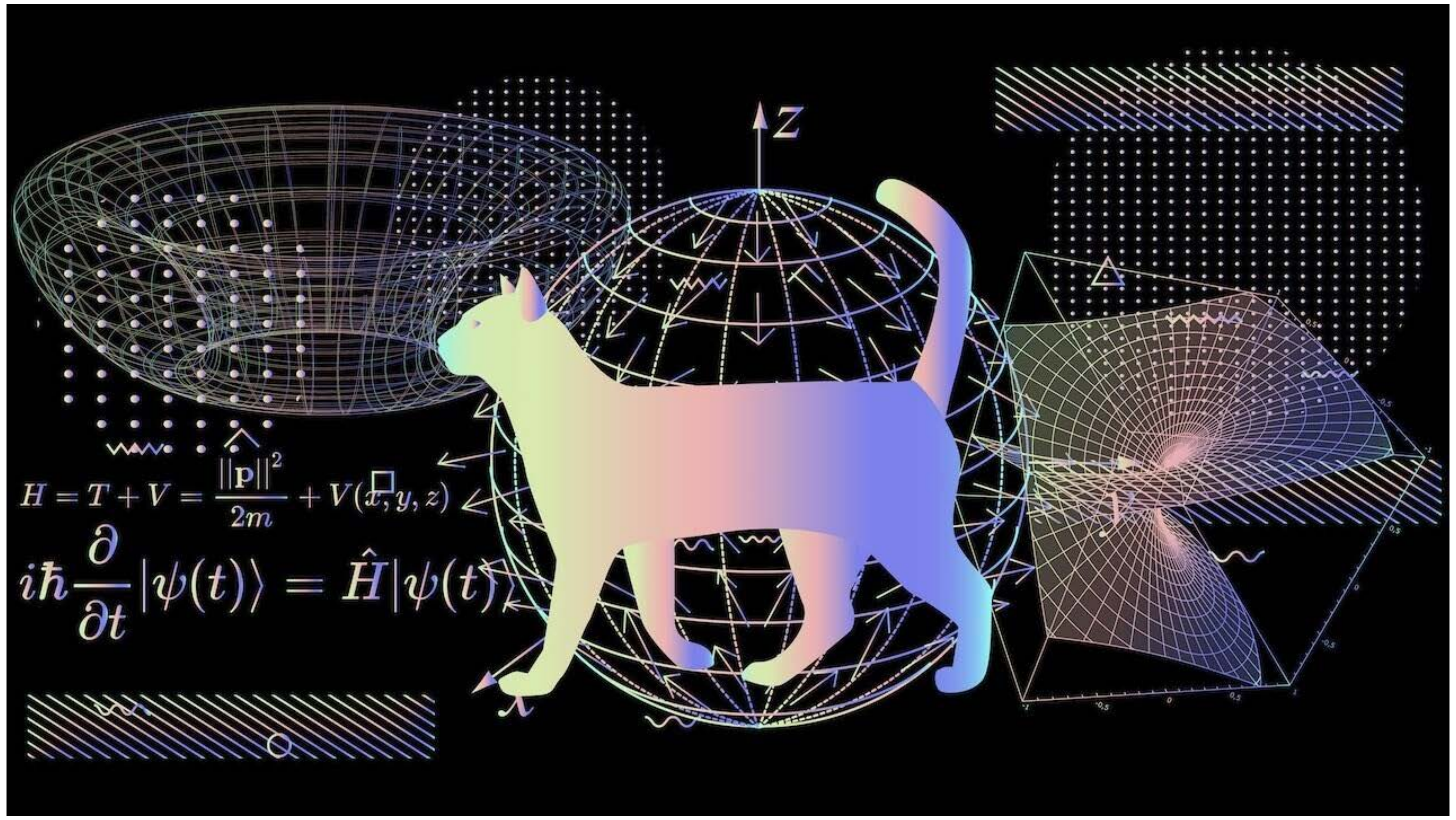
$$dP^n_{data} = |M_S + M_B|^2 dp_1 dp_2 \dots dp_n$$



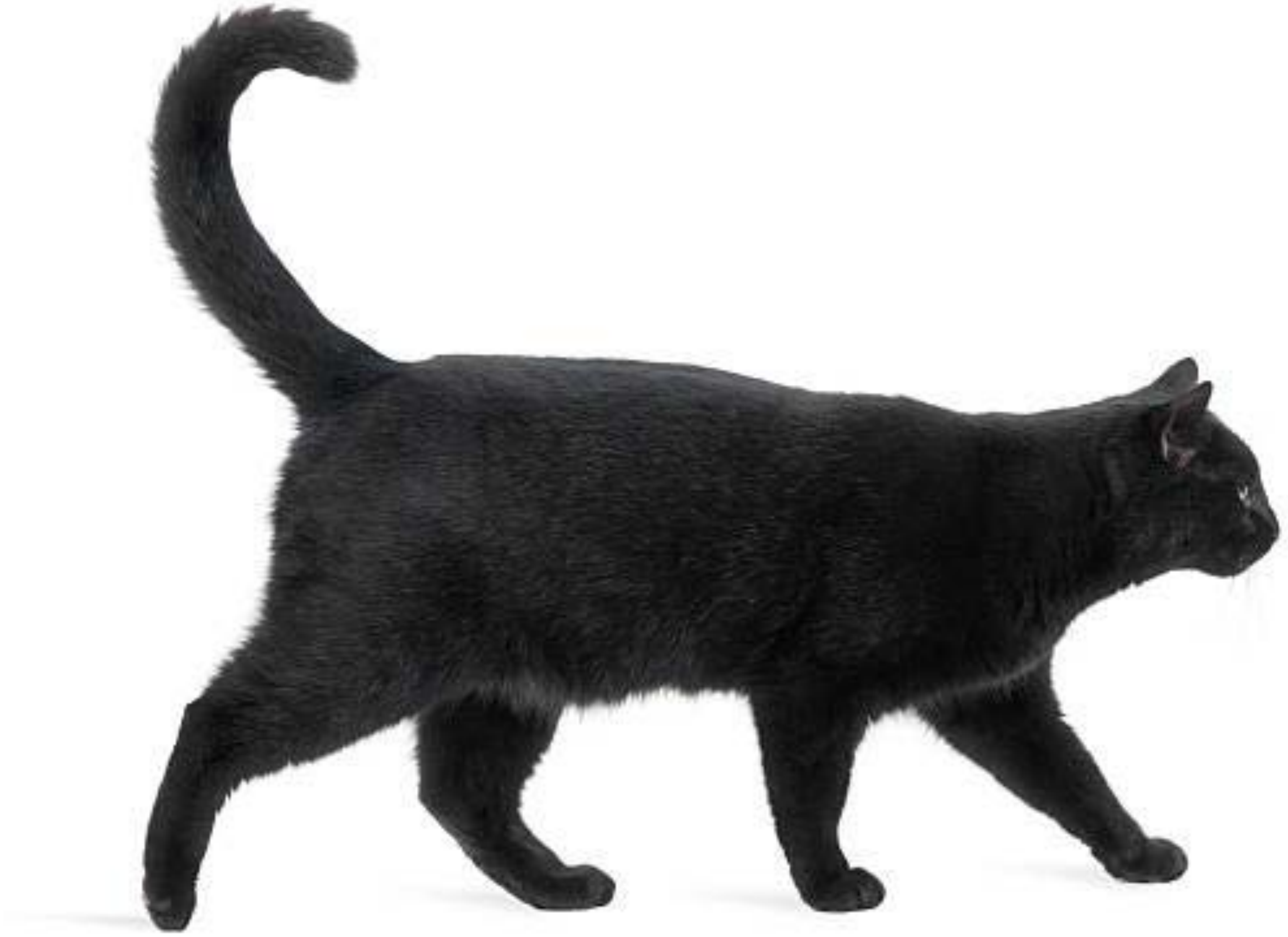
$$M_S M_B{}^* + M_B M_S{}^*$$

$$P_{data} = \alpha_S P_S + \alpha_B P_B$$




Z
$H = T + V = \frac{\|\mathbf{p}\|^2}{2m} + V(x, y, z)$
$i\hbar \frac{\partial}{\partial t} |\psi(t)\rangle = \hat{H} |\psi(t)\rangle$

!=

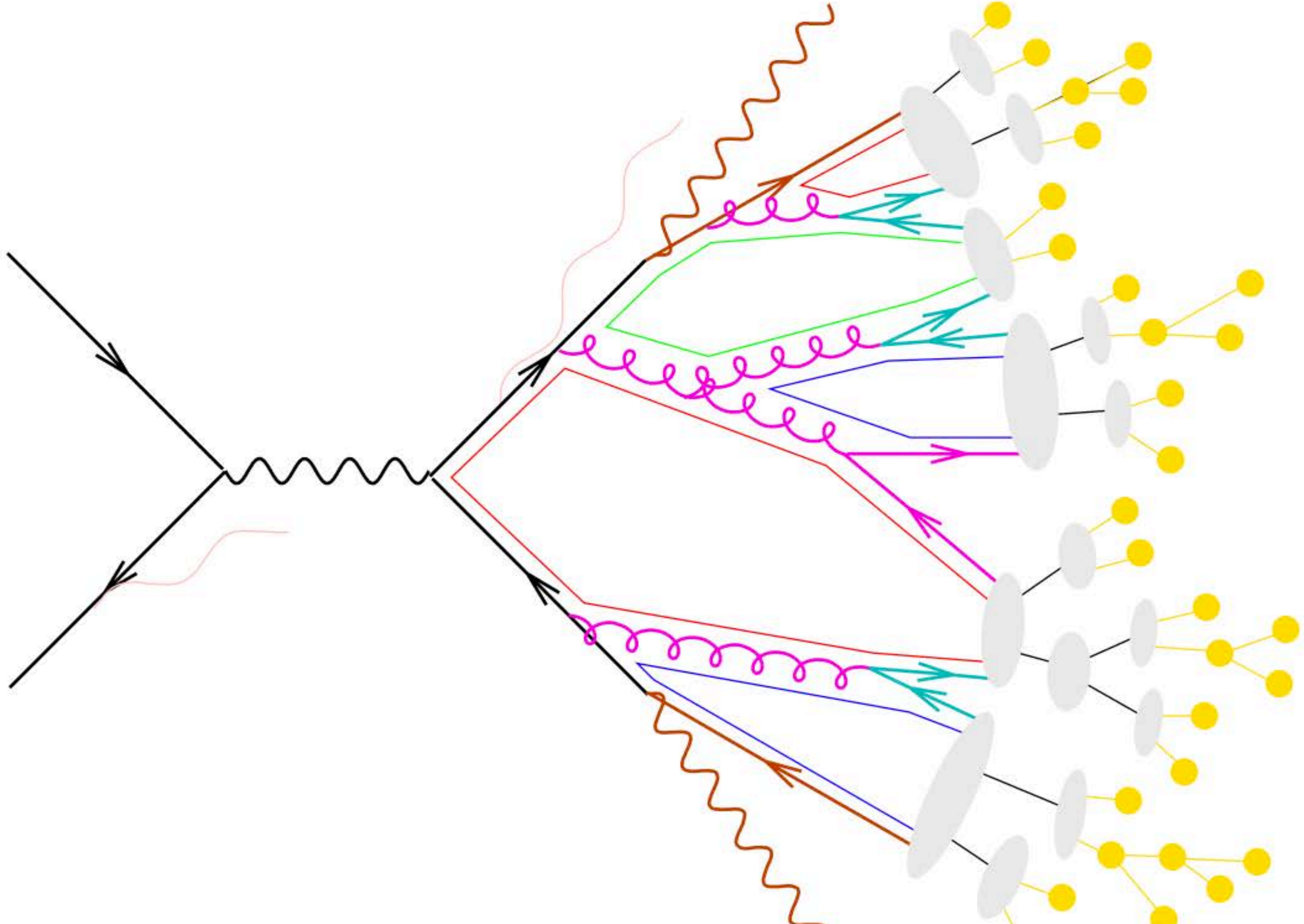


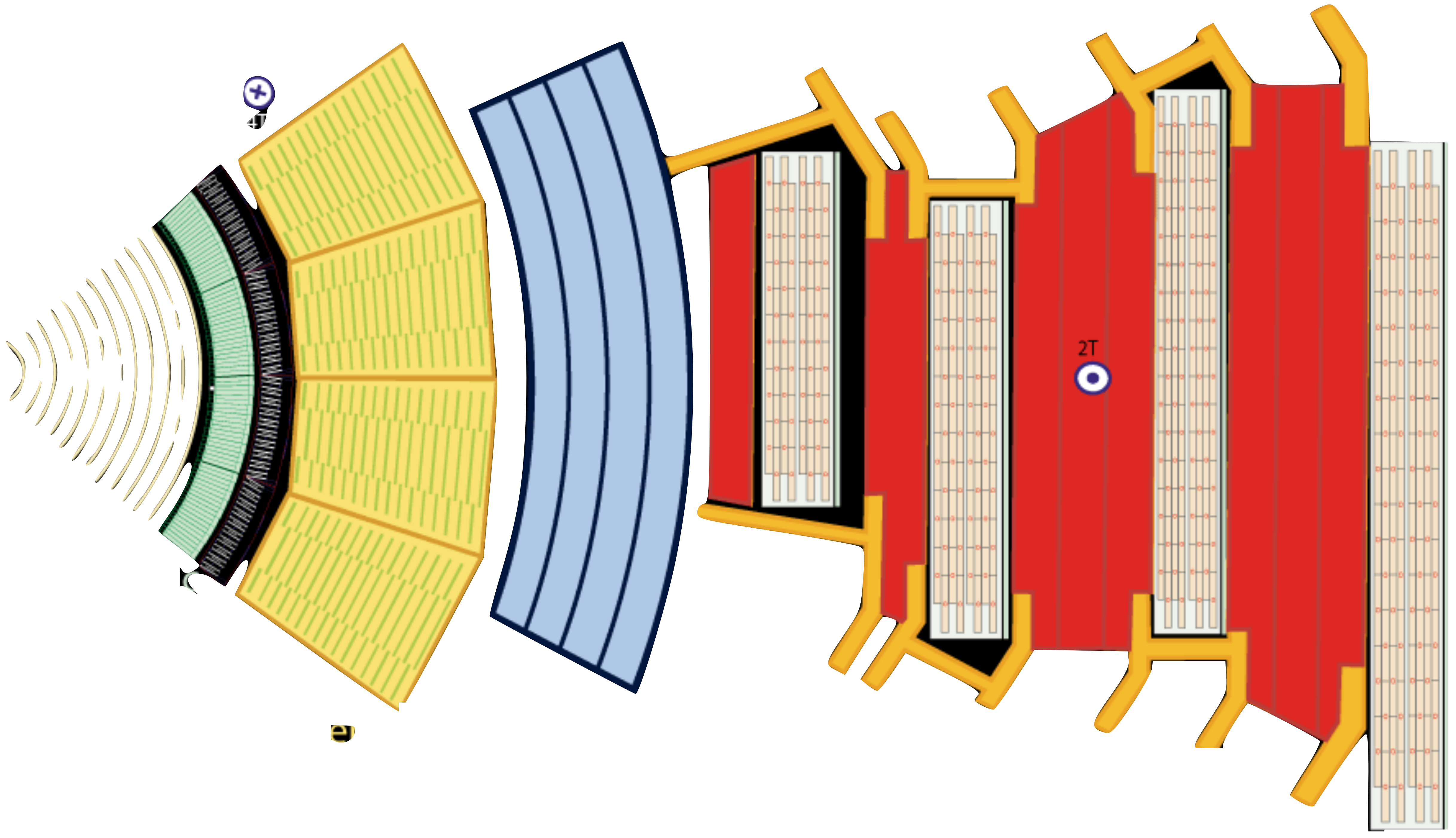
Monte Carlo Simulation

Dimensions

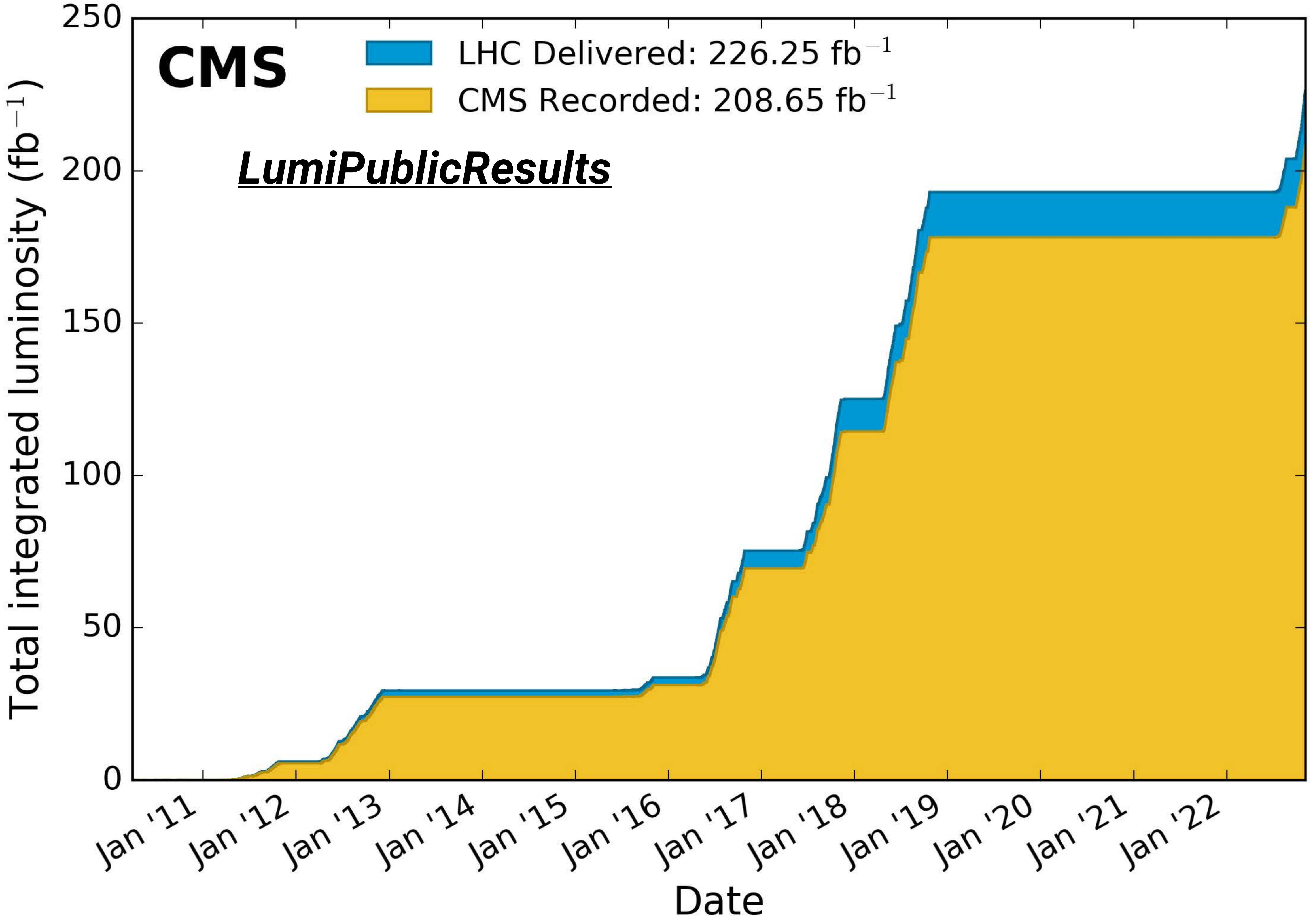
$\mathcal{O}(10)$ $\mathcal{O}(10^3)$ $\mathcal{O}(10^{10})$

10^{-18}m 10^{-15}m 10^{-6}m 100m

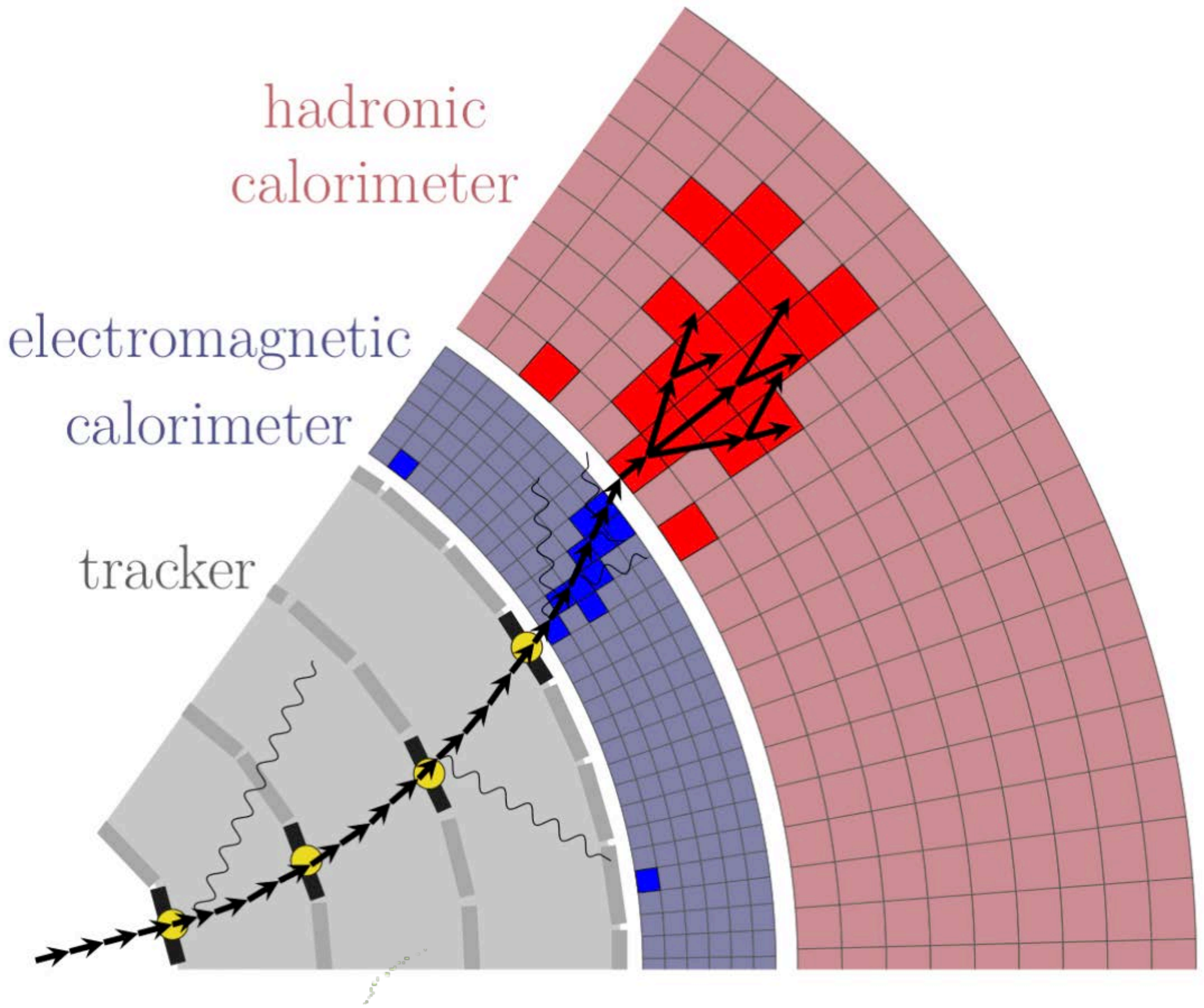


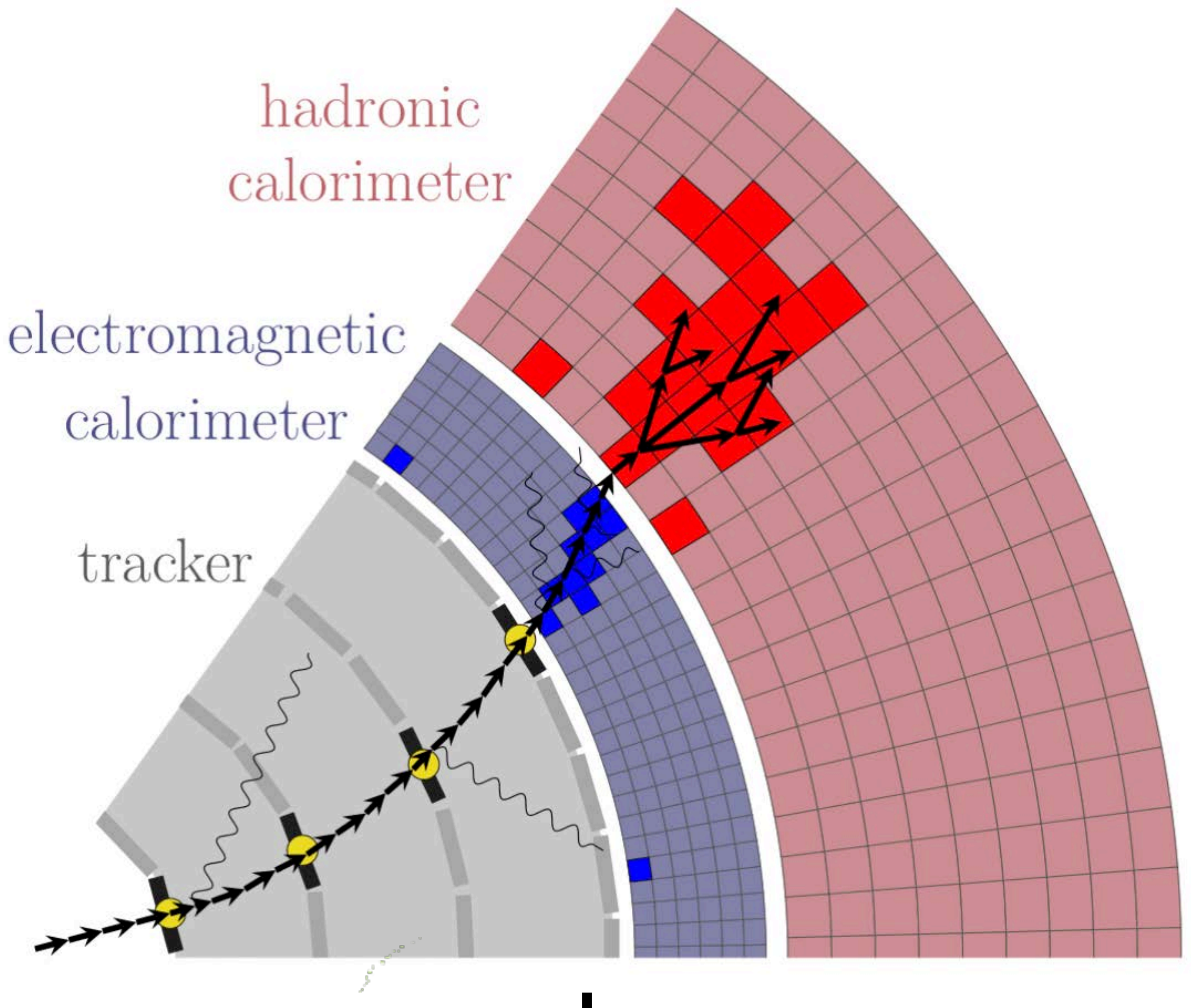


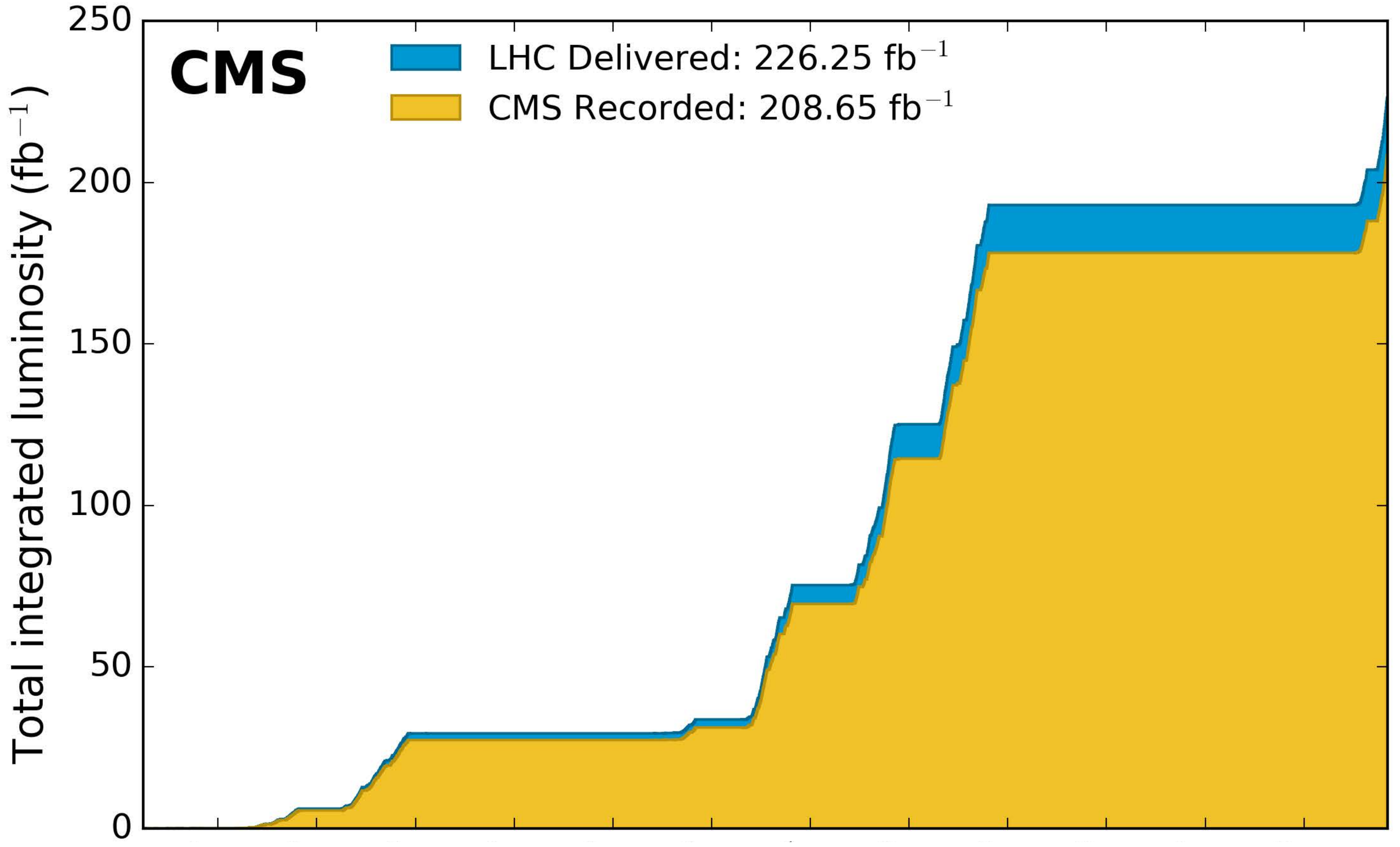
~40 quadrillion collisions recorded at LHC



O(1) trillion simulated events







Fully supervised

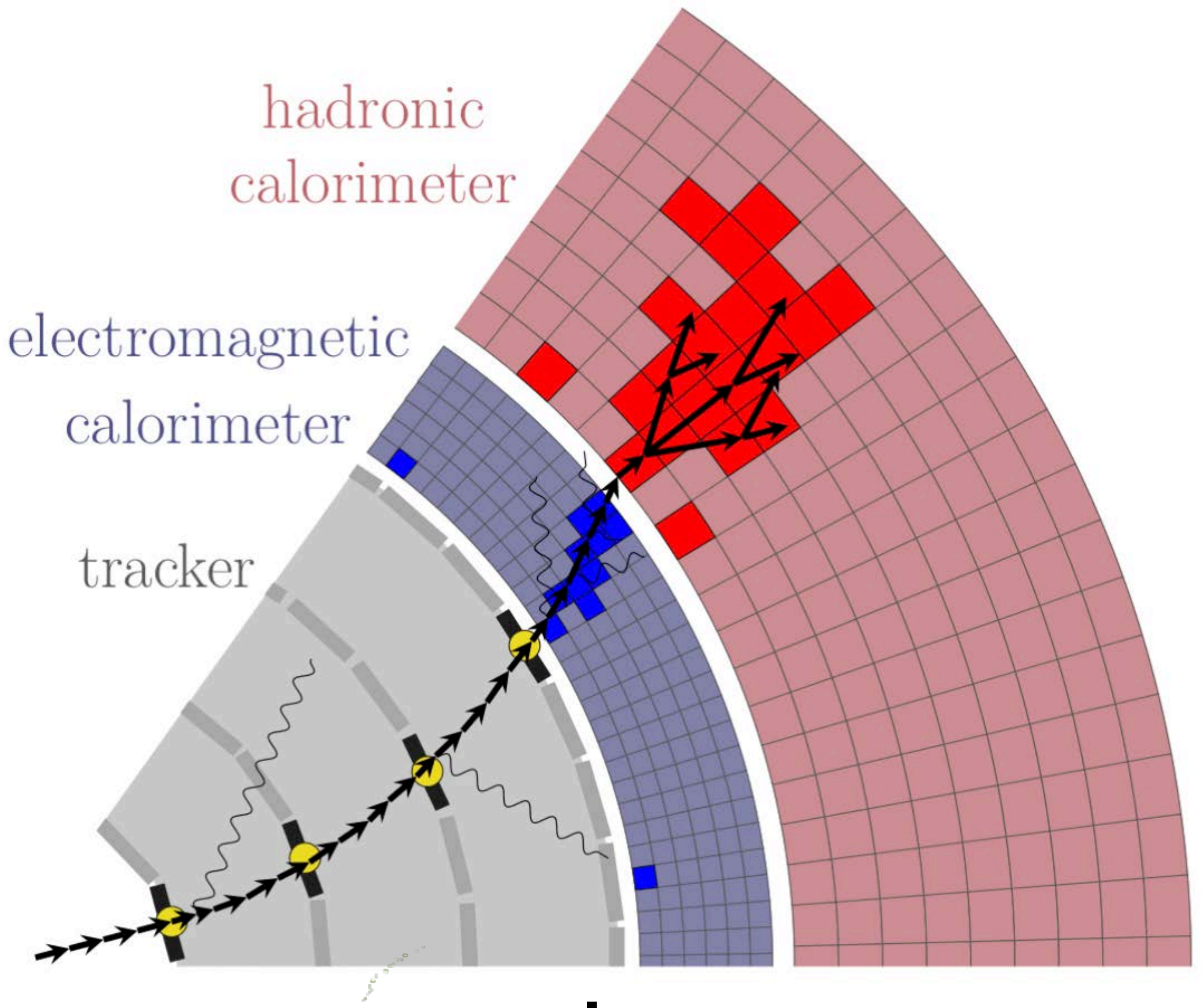
- Requires truth labels
- Only possible using simulation

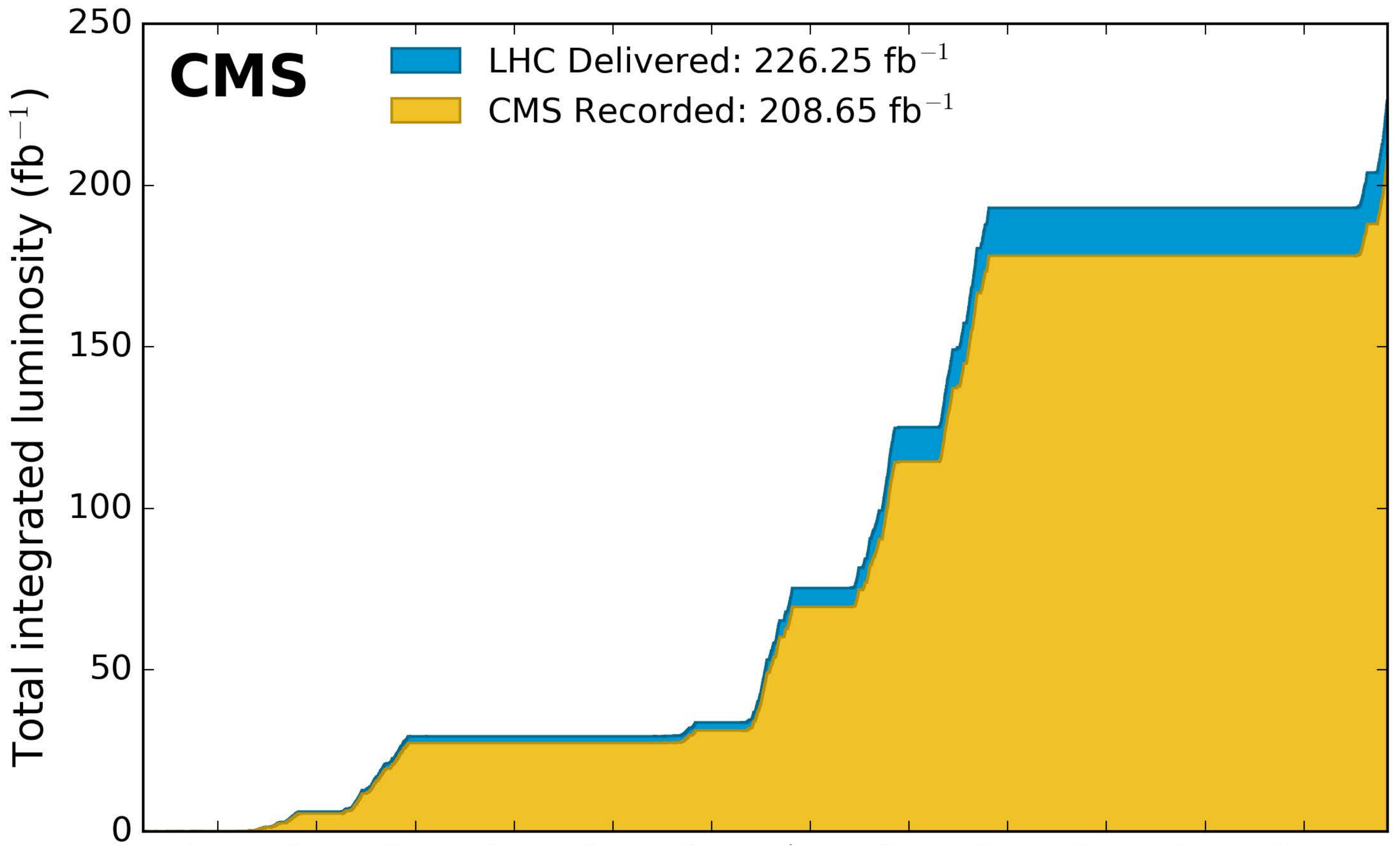
Unsupervised/SSL

No labels, completely data driven

We have a lot of high quality simulated data that we want to use

We are also very keen on using this!





Fully supervised

- Requires truth labels
- Only possible using simulation

Simulation != test data

Mostly (SM)background samples, small signal datasets

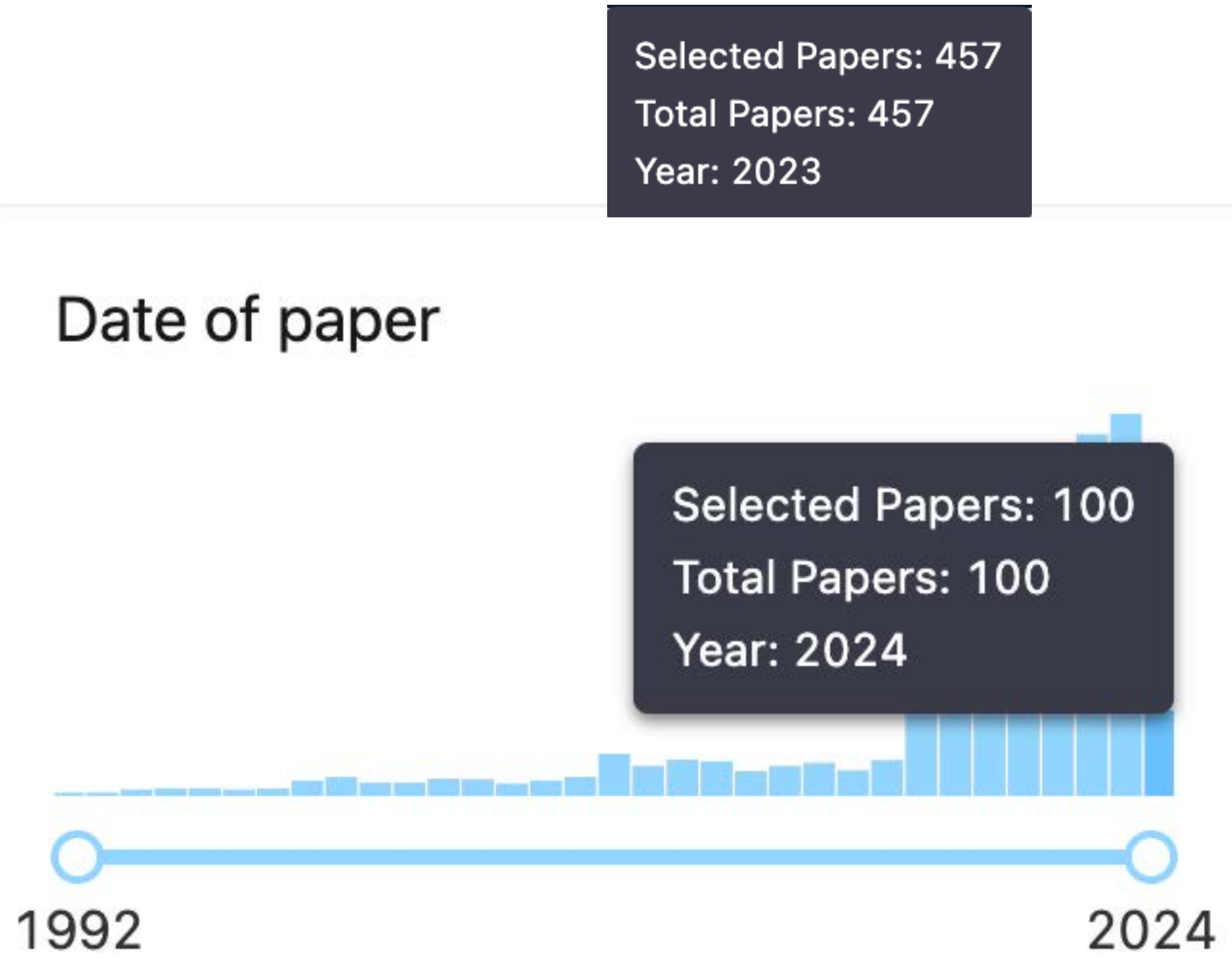
Unsupervised/SSL

No labels, completely data driven

We have a lot of high quality simulated data that we want to use

We are also very keen on using this!

Inspire:
("machine learning" or "deep learning" or neural) and (hep-ex or hep-ph or hep-th)



ROOT - An Object-Oriented Data Analysis Framework.
Authors: René Brun and Fons Rademakers
Proceedings AIHENP'96 Workshop, Lausanne, Sep. 1996, Nucl. Inst. & Meth. in Phys. Res. A 389 (1997) 81-86. See also https://root.cern/,
Date: 11th April 1997
doi: 10.1016/S0168-9002(97)00048-X
www: https://root.cern/download/lj.ps.gz
Note: Paper published in the Linux Journal, Issue 51, July 1998.

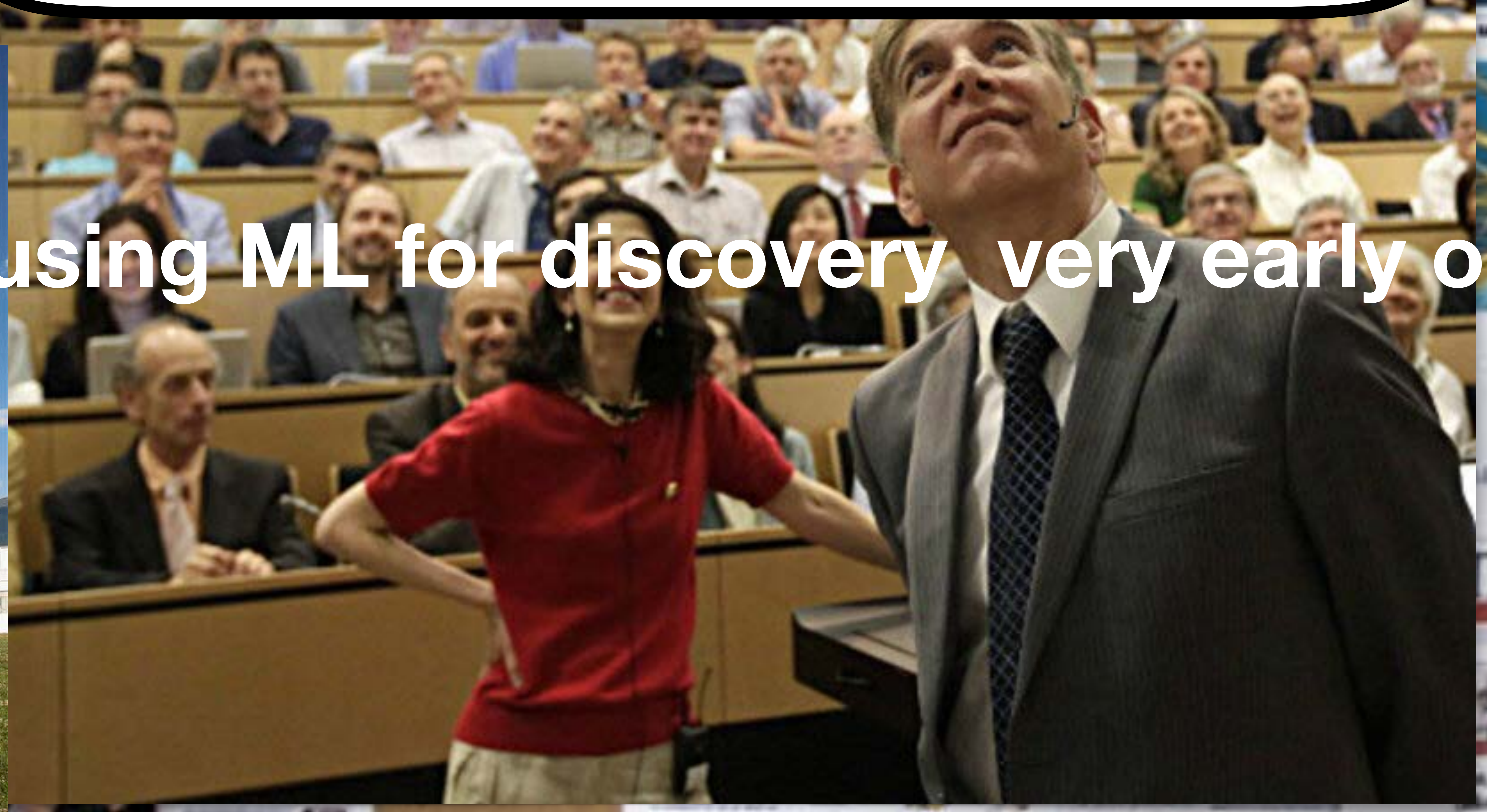
Spain bank debacle becomes legal tangle
DIE WELT
Union will Überschüsse in Sozialkassen horten
THE HINDU
“Perché l'Italia
19.7 fb^{-1} (8 TeV) + 5.1 fb^{-1} (7 TeV)
CMS
H → γγ
$m_H = \sqrt{2E_{\gamma_1}E_{\gamma_2}(1-\cos\theta_{\gamma_1\gamma_2})}$
arxiv:1407.0558
S/(S+B) weighted events / GeV
×10^3
5.7σ
B component subtracted
$m_{\gamma\gamma}$ (GeV)
ПОСЛЕДНИЙ КИРПИЧ В СТЕНУ МИРОЗДАНИЯ
МЕТРО СПУСТЯТ НА ВОДУ
'God particle' discovery has scientists giddy. Page A2
The Gazette
THE JERUSALEM POST
Mofaz rejects
CERN Summer student 2012

Nature Review

Analysis	Years of data collection	Sensitivity without machine learning	Sensitivity with machine learning	Ratio of P values	Additional data required
CMS[24] $H \to \gamma\gamma$	2011–2012	2.2σ, $P = 0.014$	2.7σ, $P = 0.0035$	4.0	51%
ATLAS[43] $H \to \tau^+\tau^-$	2011–2012	2.5σ, $P = 0.0062$	3.4σ, $P = 0.00034$	18	85%
ATLAS[99] $VH \to bb$	2011–2012	1.9σ, $P = 0.029$	2.5σ, $P = 0.0062$	4.7	73%
ATLAS[41] $VH \to bb$	2015–2016	2.8σ, $P = 0.0026$	3.0σ, $P = 0.00135$	1.9	15%
CMS[100] $VH \to bb$	2011–2012	1.4σ, $P = 0.081$	2.1σ, $P = 0.018$	4.5	125%

We were using ML for discovery very early on

CERN Summer student 2012



ACAT 2024

11–15 Mar 2024
Charles B. Wang Center, Stony Brook University
US/Eastern timezone

Enter your search term

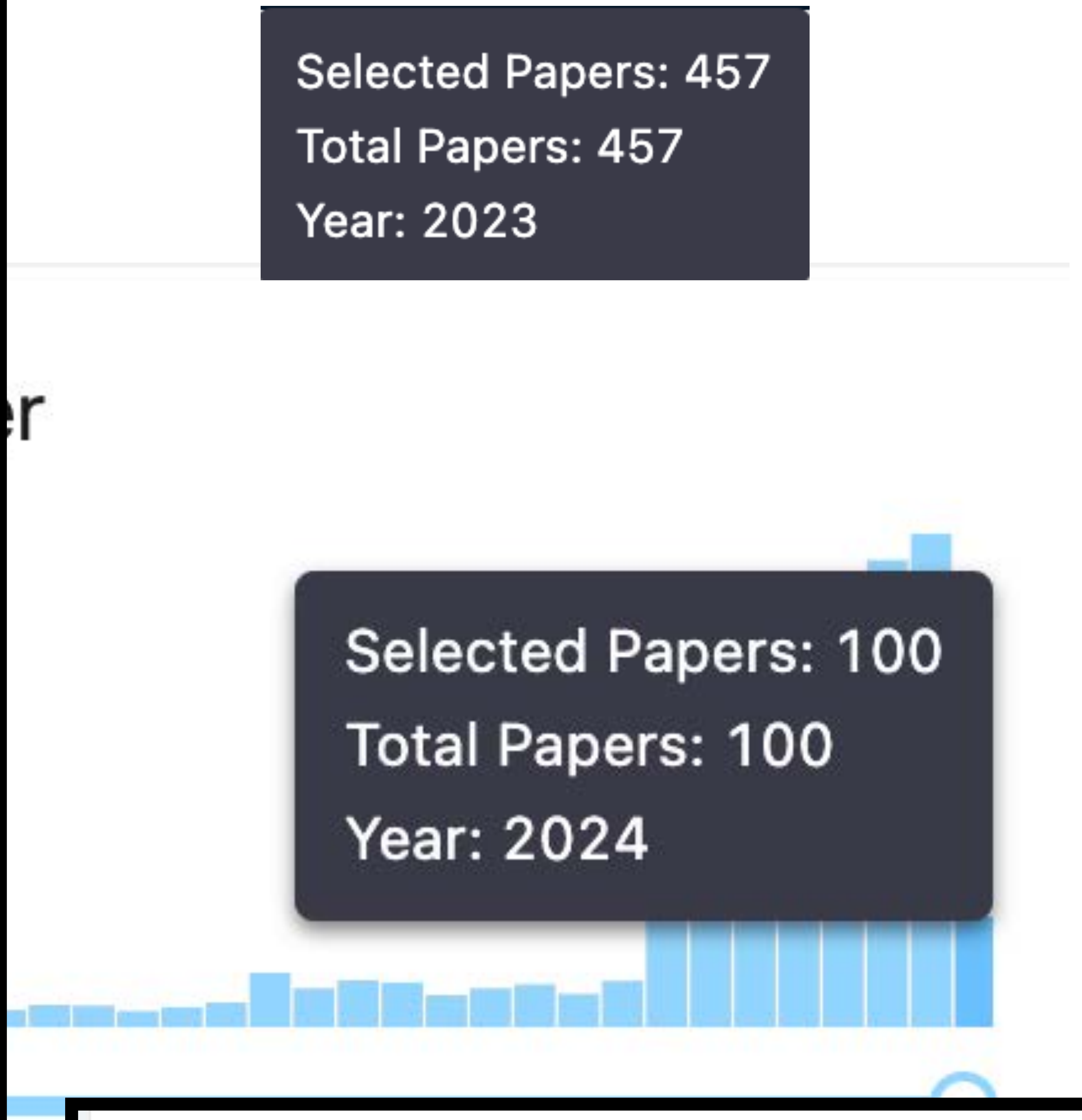
Now happening: Exascale infrastructures for Science (Theatre) *08:45 - 09:15*

22nd International Workshop on Advanced Computing and Analysis Techniques in Physics Research

The 22nd International Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT 2024 will take place between Monday 11th and Friday, 15th March, 2024 at the Stony Brook University, Stony Brook, Long Island NY, USA.

The 22nd edition of ACAT will — once again — bring together computational experts from a wide range of disciplines, including particle-, nuclear-, astro-, and accelerator-physics as well as high performance computing. Through this unique forum, we will explore the areas where these disciplines overlap with computer science, fostering the exchange of ideas related to cutting-edge computing, data-analysis, and theoretical-calculation technologies.

Our Theme will be **Foundation Models for Physics - Nexus of Computation and Physics through Embracing the Era of Foundation Models:** The 2024 ACAT workshop invites the vanguard of computational and physics experts to delve into the transformative potential of foundation models. As the intersection between physics and computational realms deepens, these advanced models, underpinned by colossal datasets and capable of generating nuanced outputs, are redefining the research spectrum and increasingly reshaping the way researchers approach complex problems, simulations, and data analyses. As we chart this new territory, we'll address challenges and opportunities encompassing integration into computational ecosystems, innovative data practices, training nuances, infrastructure evolution, uncertainty metrics, ethical dimensions, and collaborative vistas across disciplines.



Re-Simulation-based Self-Supervised Learning for Pre-Training Foundation Models #3
Philip Harris, Michael Kagan, Jeffrey Krupa, Benedikt Maier, Nathaniel Woodward (Mar 11, 2024)
e-Print: 2403.07066 [hep-ph]
pdf cite claim reference search 0 citations

OmniJet-α: The first cross-task foundation model for particle physics #5
Joschka Birk, Anna Hallin, Gregor Kasieczka (Mar 8, 2024)
e-Print: 2403.05618 [hep-ph]
pdf cite claim reference search 0 citations

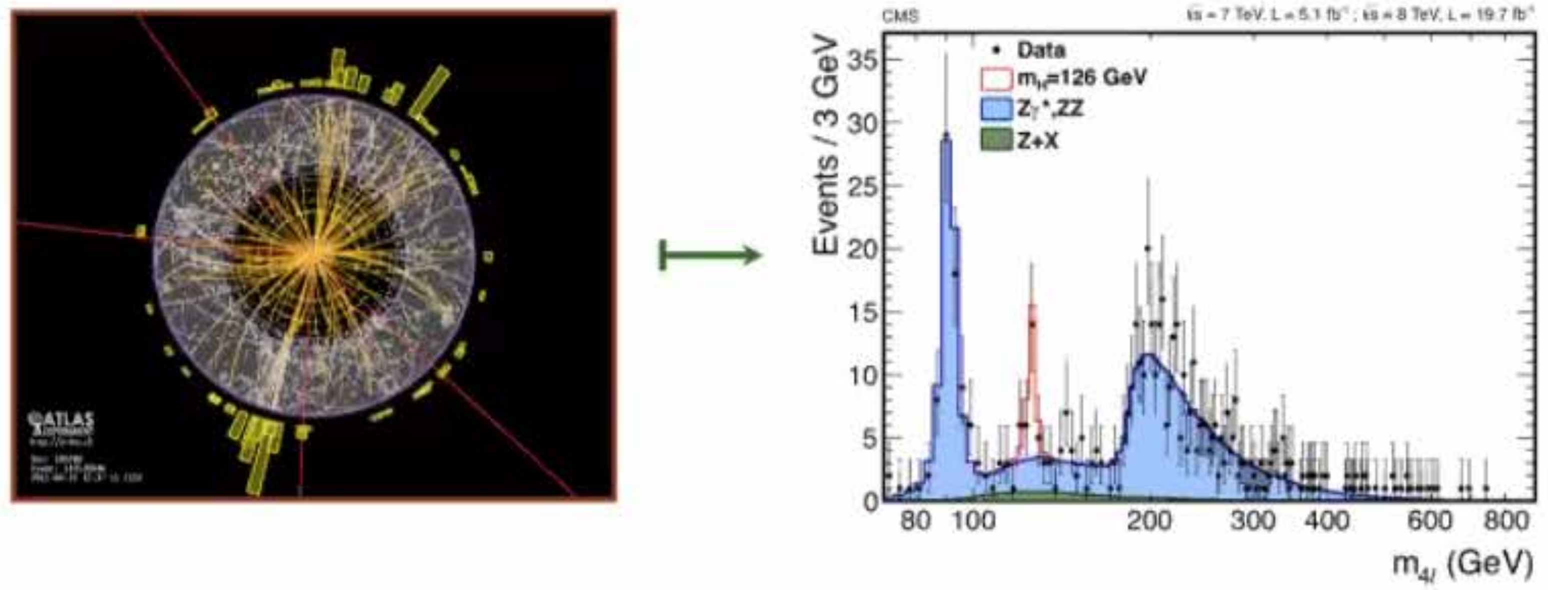
From Siddhartha’s introduction

Framing: Kyle Cranmer

AI + Physics: *A new frontier?*

Many fields within AI4Science are pushing the frontiers of AI… what about physics?

Reliable inference with complex forward models



- Sampling under complex symmetries and exactness guarantees (e.g., in lattice QFT)
- Statistical anomaly detection
- Highly structured models/data-generating processes
- …

Extremely fast real-time inference

High Energy Physics

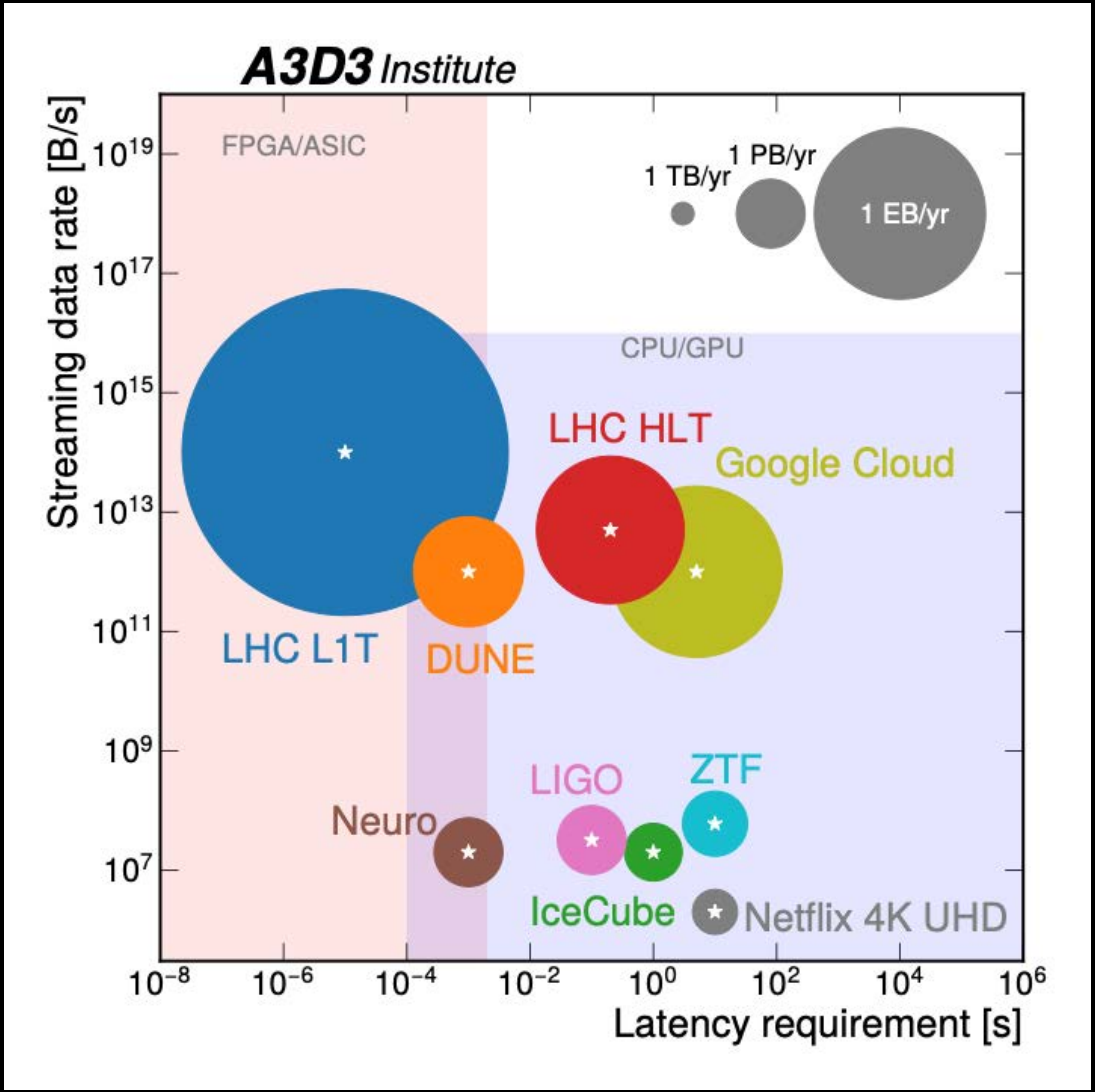
Build tools to process LHC collisions occurring 40 million times per second data in real-time using AI.

Read More >

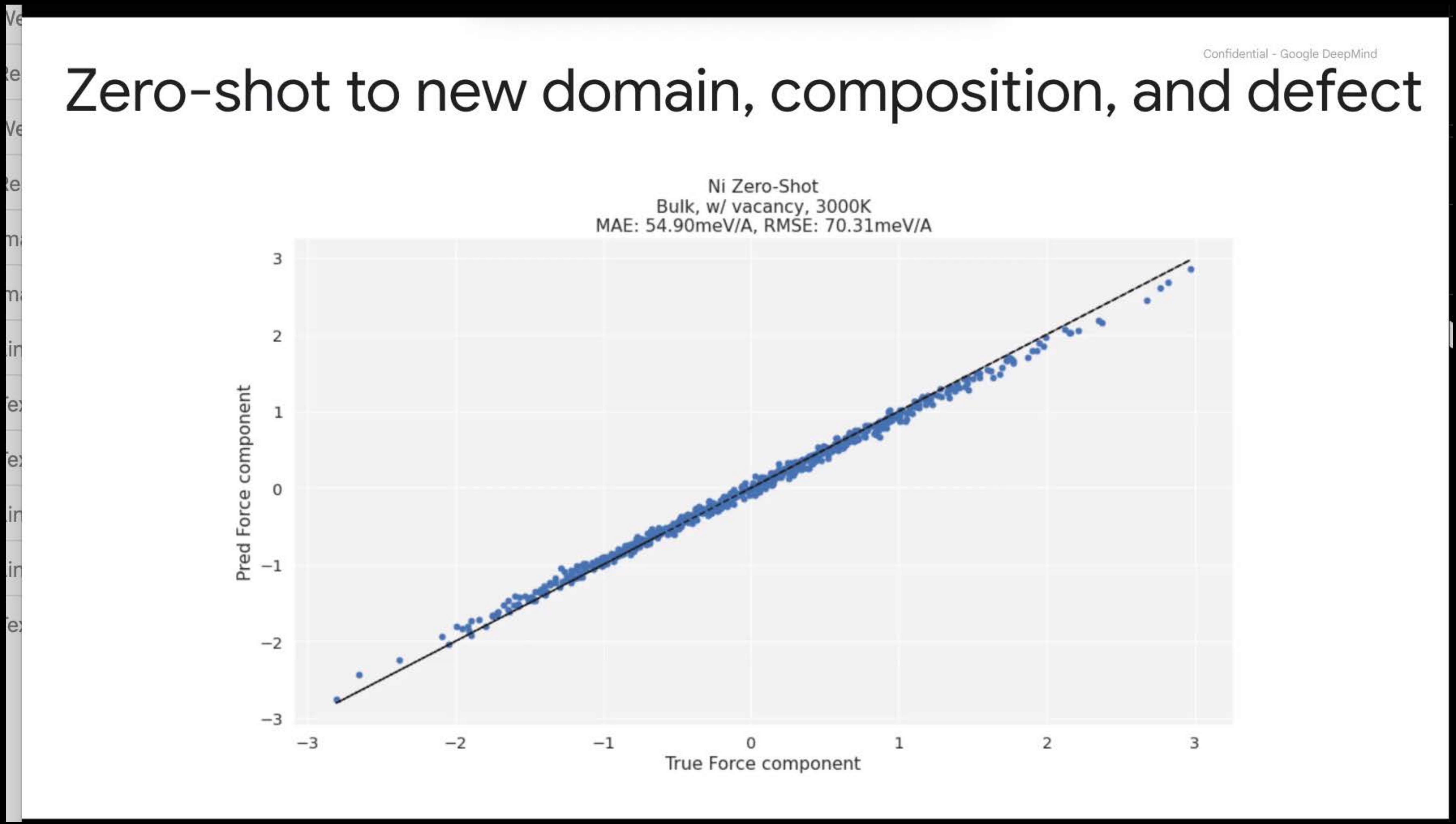
(From A3D3 website)

FastML:

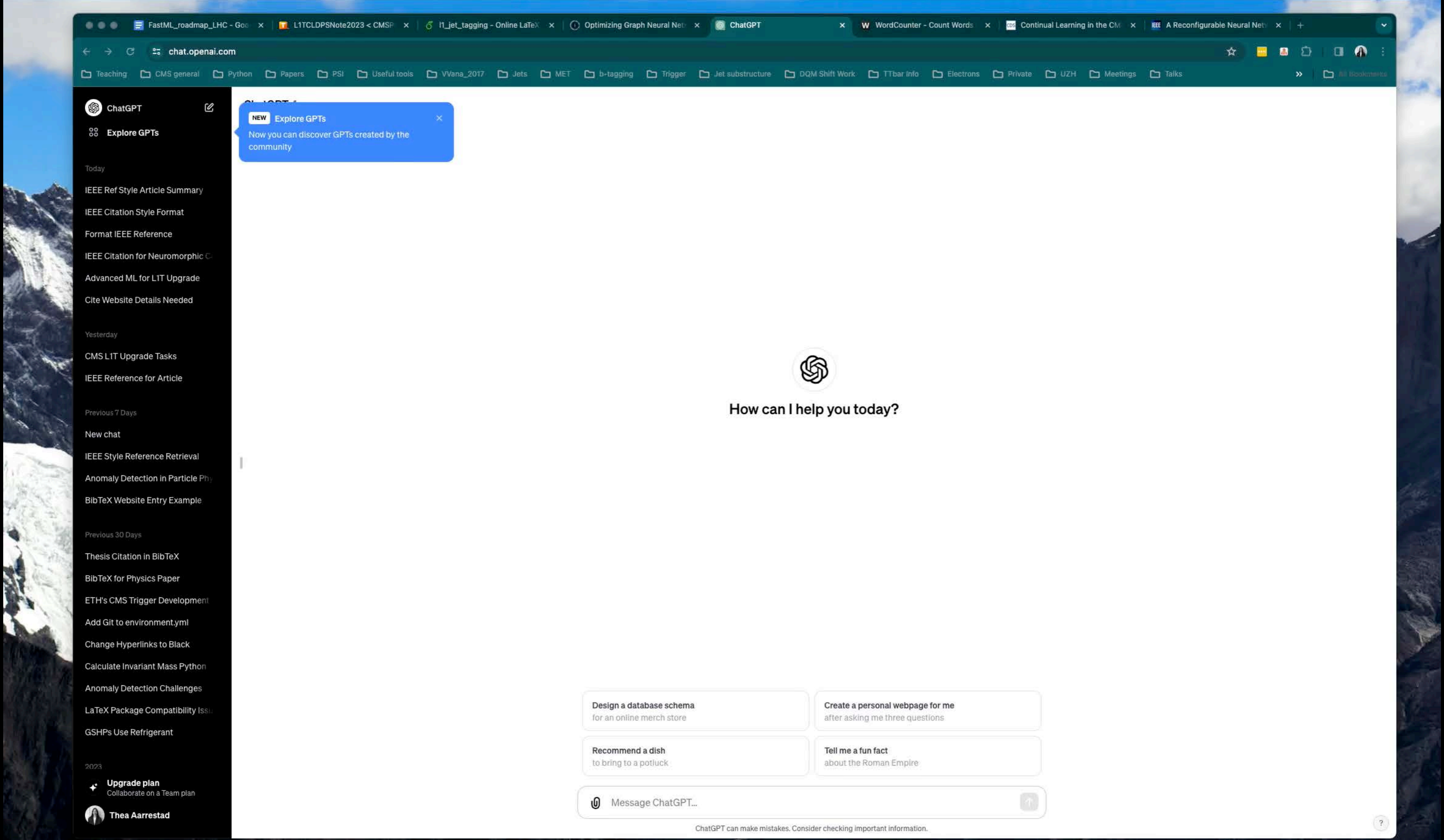
Pioneering AI in the physical sciences

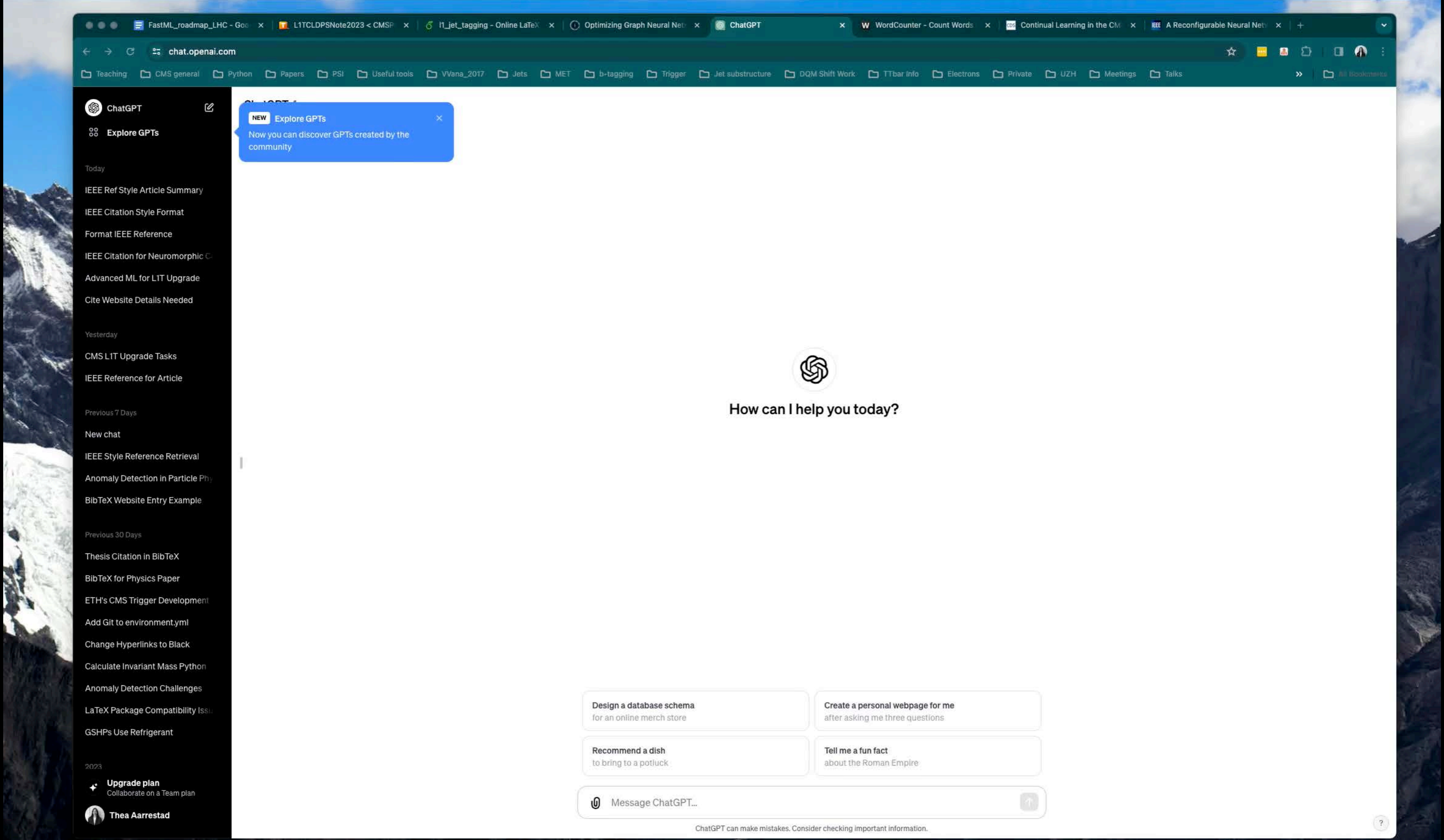


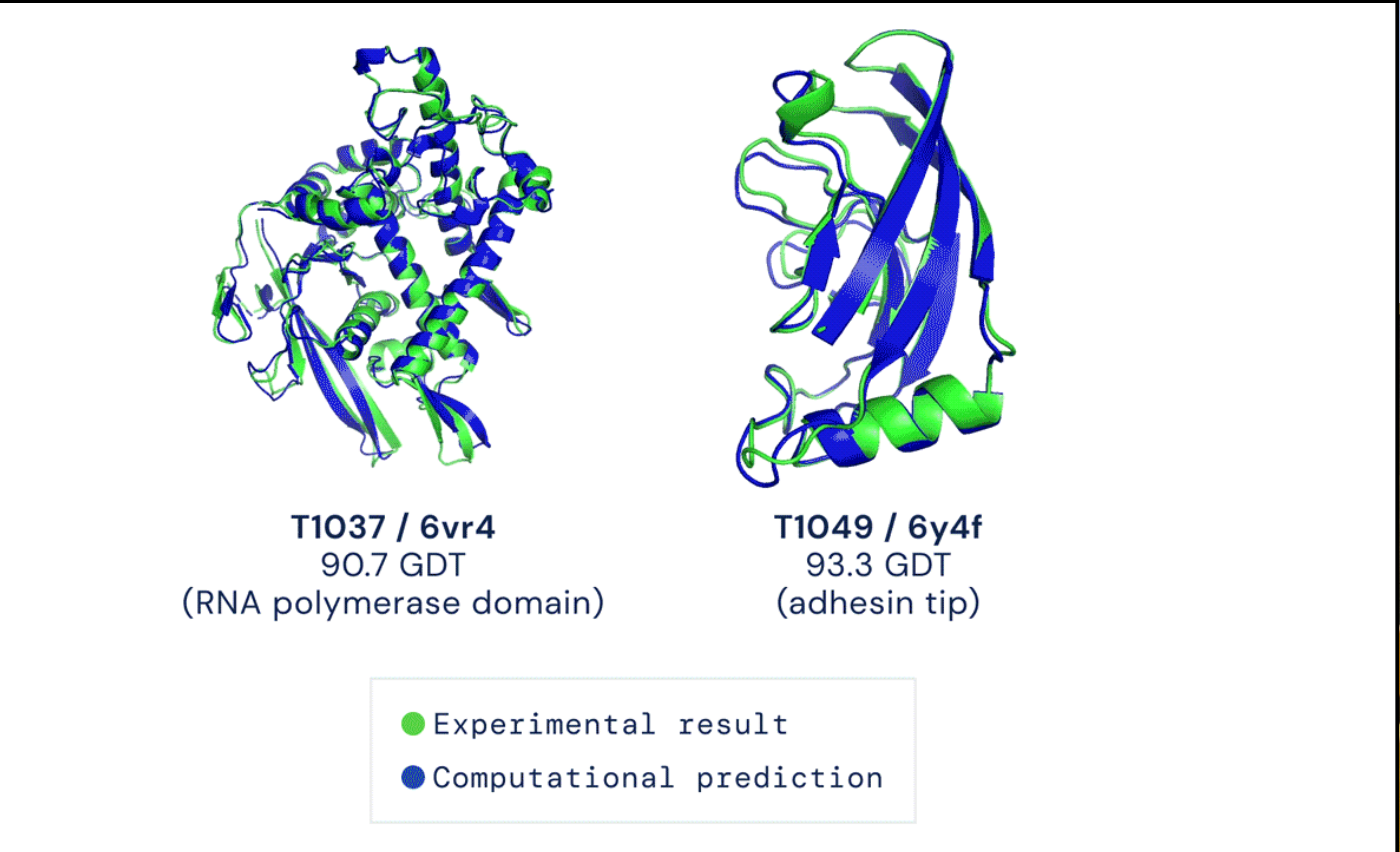
From Simon. 60 million parameter model



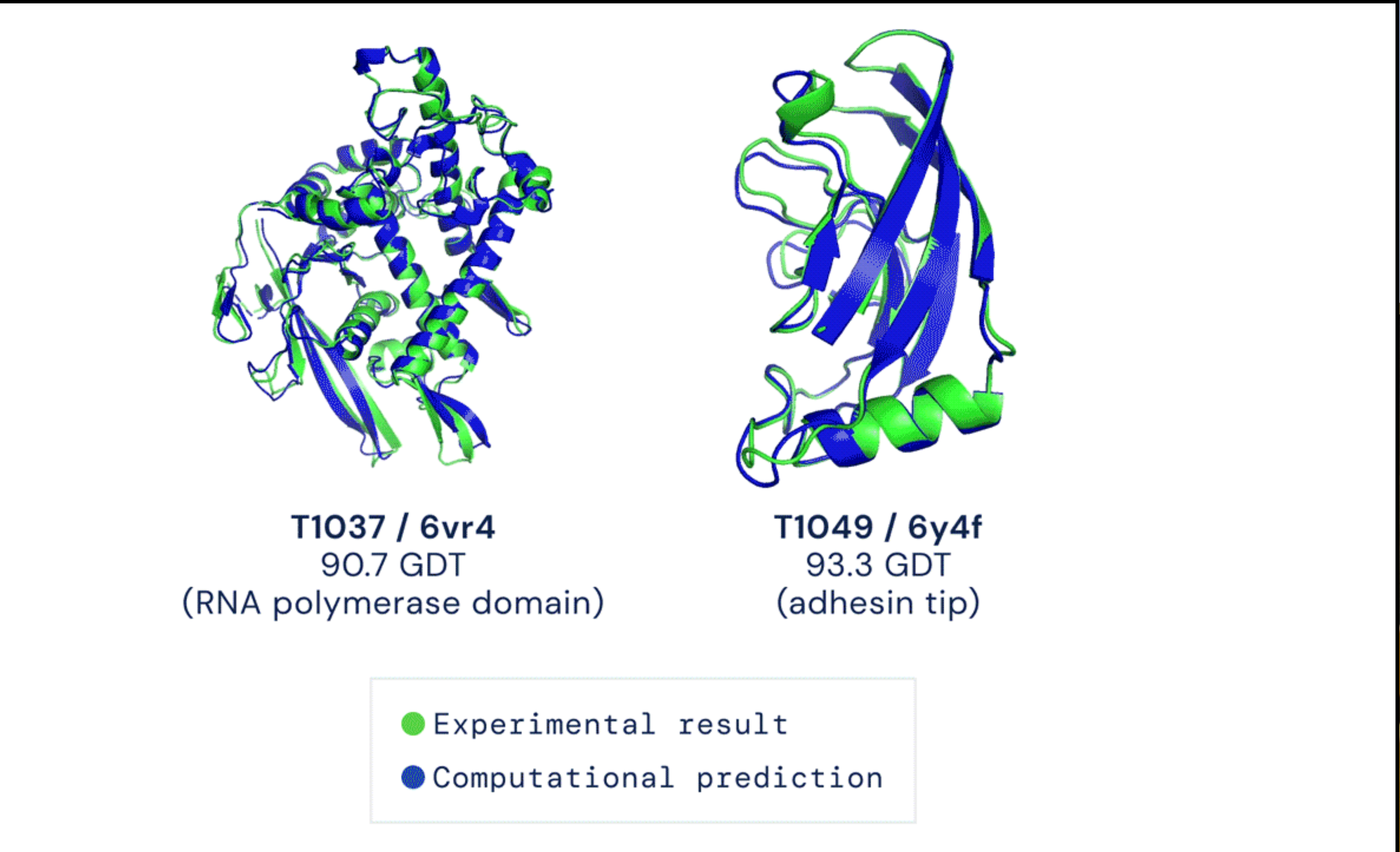
Can we combine 12μs latency and O(100M) parameter models?

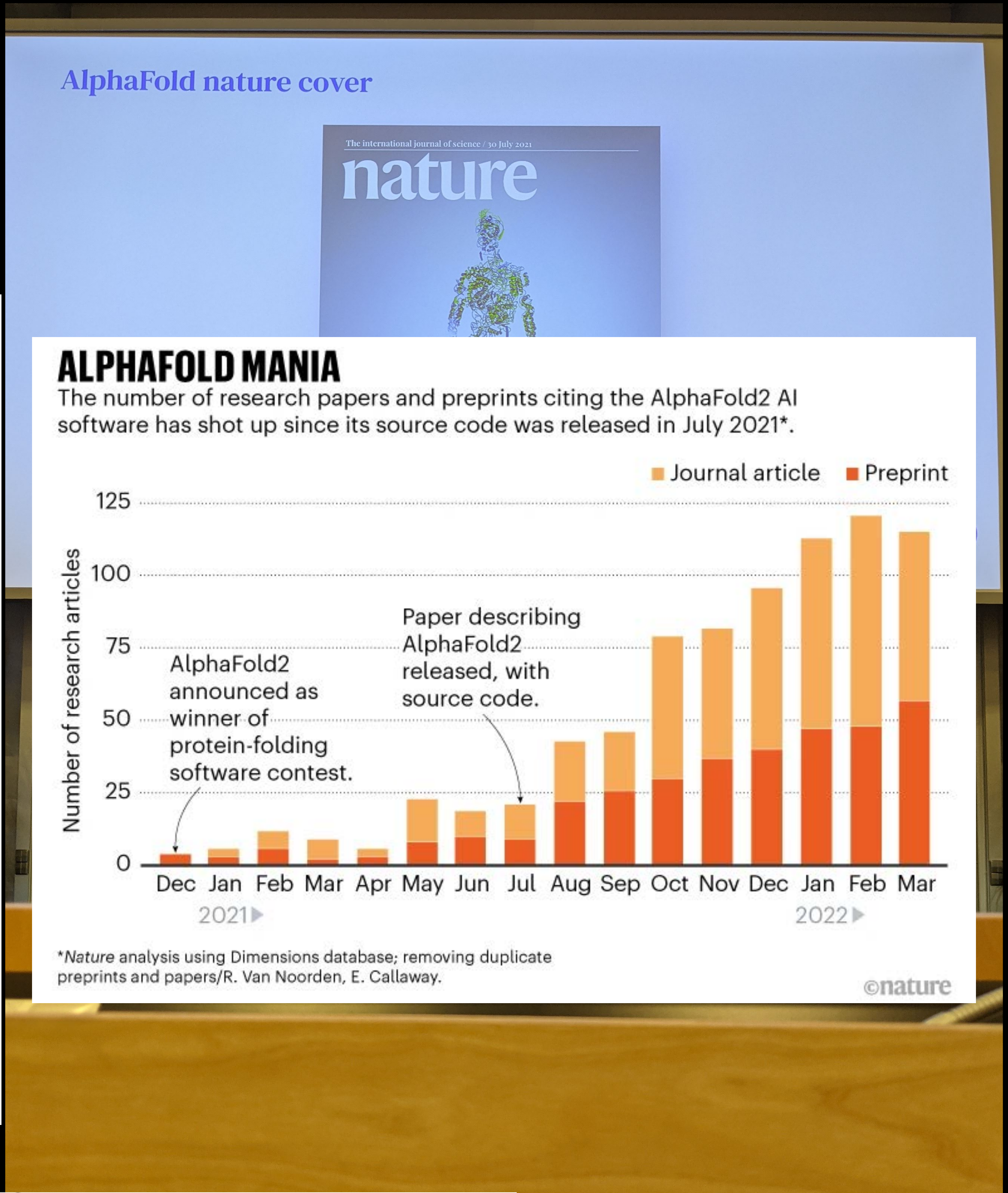

FastML_roadmap_LHC - Goo
L1TCLDPSNote2023 < CMSP
l1_jet_tagging - Online LaTeX
Optimizing Graph Neural Net
ChatGPT
WordCounter - Count Words
Continual Learning in the CM
A Reconfigurable Neural Net
chat.openai.com
Teaching
CMS general
Python
Papers
PSI
Useful tools
VVana_2017
Jets
MET
b-tagging
Trigger
Jet substructure
DQM Shift Work
TTbar info
Electrons
Private
UZH
Meetings
Talks
ChatGPT
Explore GPTs
NEW
Explore GPTs
Now you can discover GPTs created by the community
Today
IEEE Ref Style Article Summary
IEEE Citation Style Format
Format IEEE Reference
IEEE Citation for Neuromorphic C
Advanced ML for L1T Upgrade
Cite Website Details Needed
Yesterday
CMS L1T Upgrade Tasks
IEEE Reference for Article
Previous 7 Days
New chat
IEEE Style Reference Retrieval
Anomaly Detection in Particle Ph
BibTeX Website Entry Example
Previous 30 Days
Thesis Citation in BibTeX
BibTeX for Physics Paper
ETH's CMS Trigger Development
Add Git to environment.yml
Change Hyperlinks to Black
Calculate Invariant Mass Python
Anomaly Detection Challenges
LaTeX Package Compatibility Issu
GSHPs Use Refrigerant
2023
Upgrade plan
Collaborate on a Team plan
Thea Aarrestad
How can I help you today?
Design a database schema
for an online merch store
Create a personal webpage for me
after asking me three questions
Recommend a dish
to bring to a potluck
Tell me a fun fact
about the Roman Empire
Message ChatGPT...
ChatGPT can make mistakes. Consider checking important information.


FastML_roadmap_LHC - Goo
L1TCLDPSNote2023 < CMSP
l1_jet_tagging - Online LaTeX
Optimizing Graph Neural Net
ChatGPT
WordCounter - Count Words
Continual Learning in the CM
A Reconfigurable Neural Net
chat.openai.com
Teaching
CMS general
Python
Papers
PSI
Useful tools
VVana_2017
Jets
MET
b-tagging
Trigger
Jet substructure
DQM Shift Work
TTbar info
Electrons
Private
UZH
Meetings
Talks
ChatGPT
Explore GPTs
NEW
Explore GPTs
Now you can discover GPTs created by the community
Today
IEEE Ref Style Article Summary
IEEE Citation Style Format
Format IEEE Reference
IEEE Citation for Neuromorphic C
Advanced ML for L1T Upgrade
Cite Website Details Needed
Yesterday
CMS L1T Upgrade Tasks
IEEE Reference for Article
Previous 7 Days
New chat
IEEE Style Reference Retrieval
Anomaly Detection in Particle Ph
BibTeX Website Entry Example
Previous 30 Days
Thesis Citation in BibTeX
BibTeX for Physics Paper
ETH's CMS Trigger Development
Add Git to environment.yml
Change Hyperlinks to Black
Calculate Invariant Mass Python
Anomaly Detection Challenges
LaTeX Package Compatibility Iss
GSHPs Use Refrigerant
2023
Upgrade plan
Collaborate on a Team plan
Thea Aarrestad
How can I help you today?
Design a database schema
for an online merch store
Create a personal webpage for me
after asking me three questions
Recommend a dish
to bring to a potluck
Tell me a fun fact
about the Roman Empire
Message ChatGPT...
ChatGPT can make mistakes. Consider checking important information.



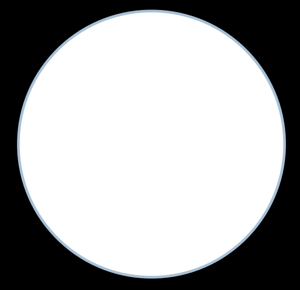
sequence—the structure prediction component of the 'protein folding problem'[8]—has been an important open research problem for more than 50 years[9]. Despite recent





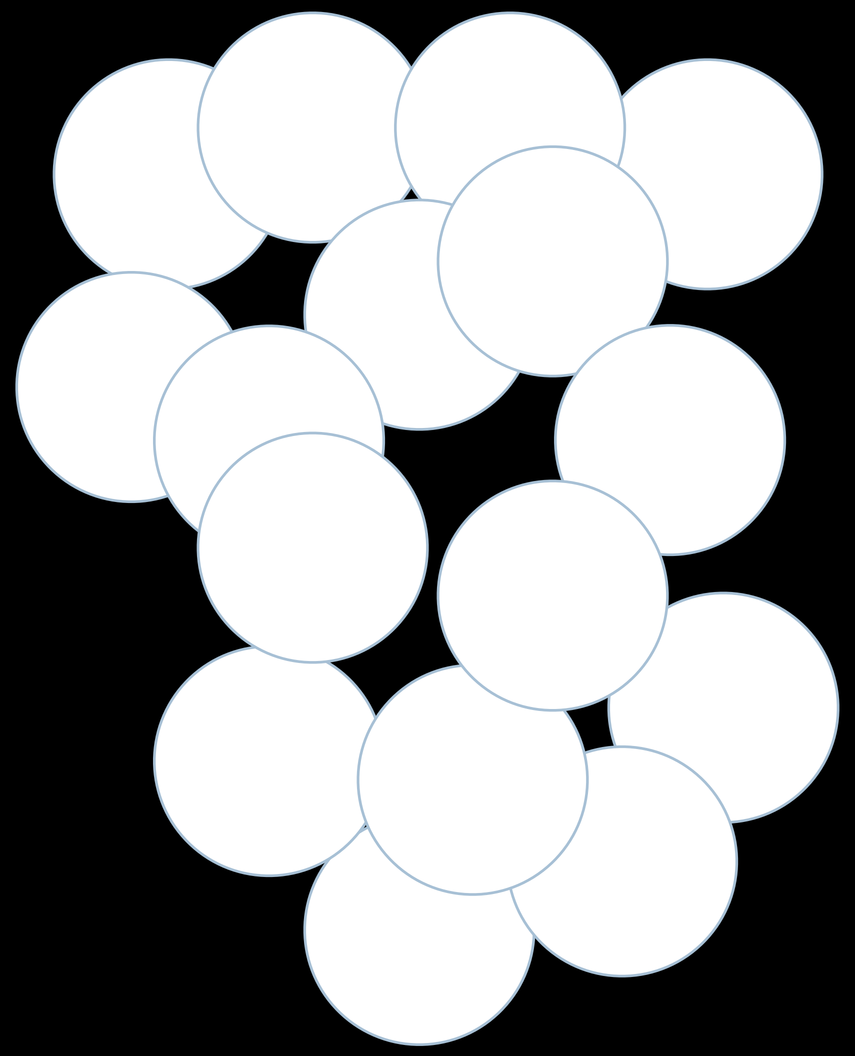
sequence—the structure prediction component of the 'protein folding problem'[8]—has been an important open research problem for more than 50 years[9]. Despite recent

GPT-3



175,000,000,000
(0.16% of neurons in your brain)

GPT-4



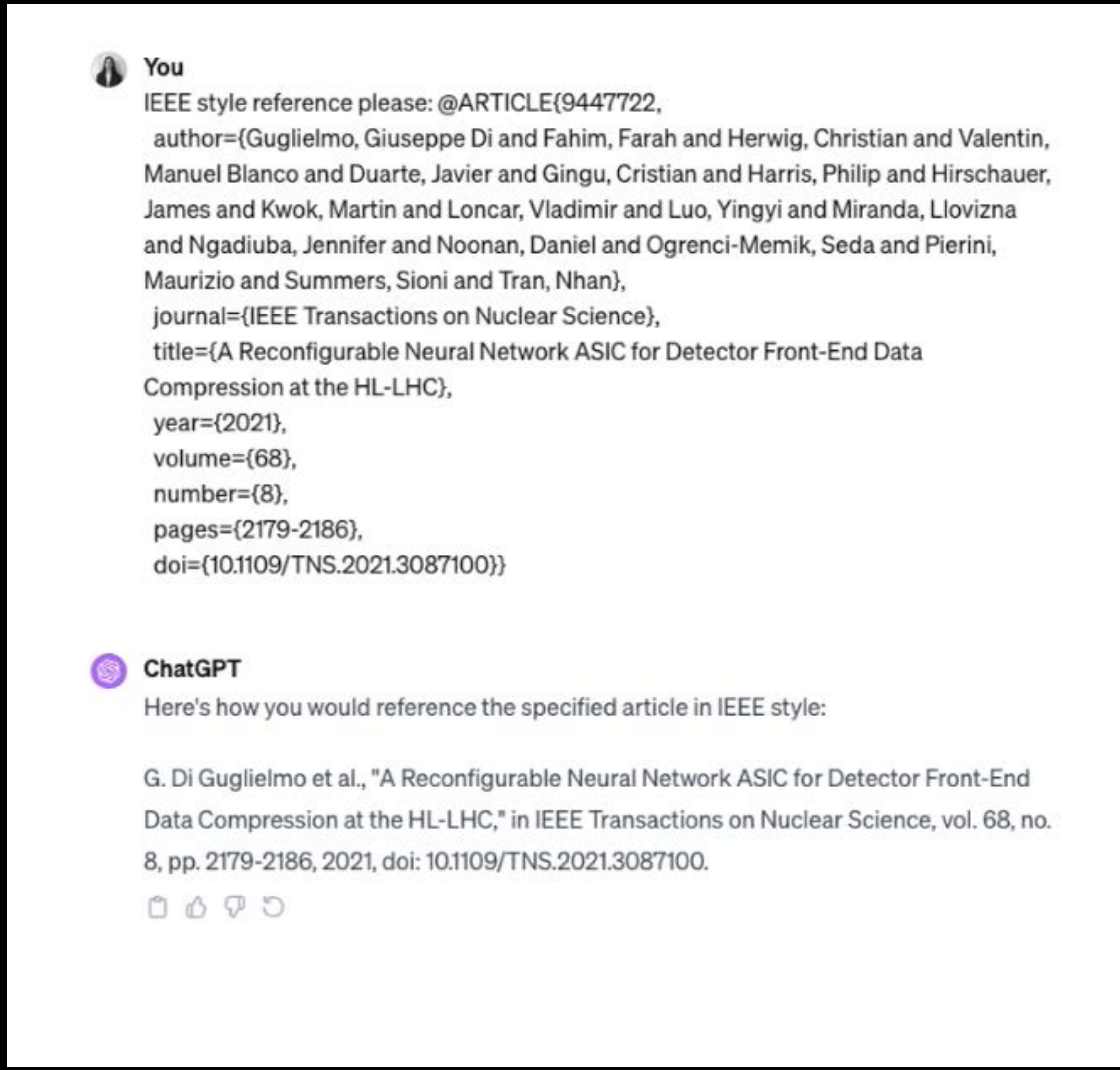
1,800,000,000,000
(1.6% of neurons in your brain)

semianalysis 2023

Train (GPT-4):

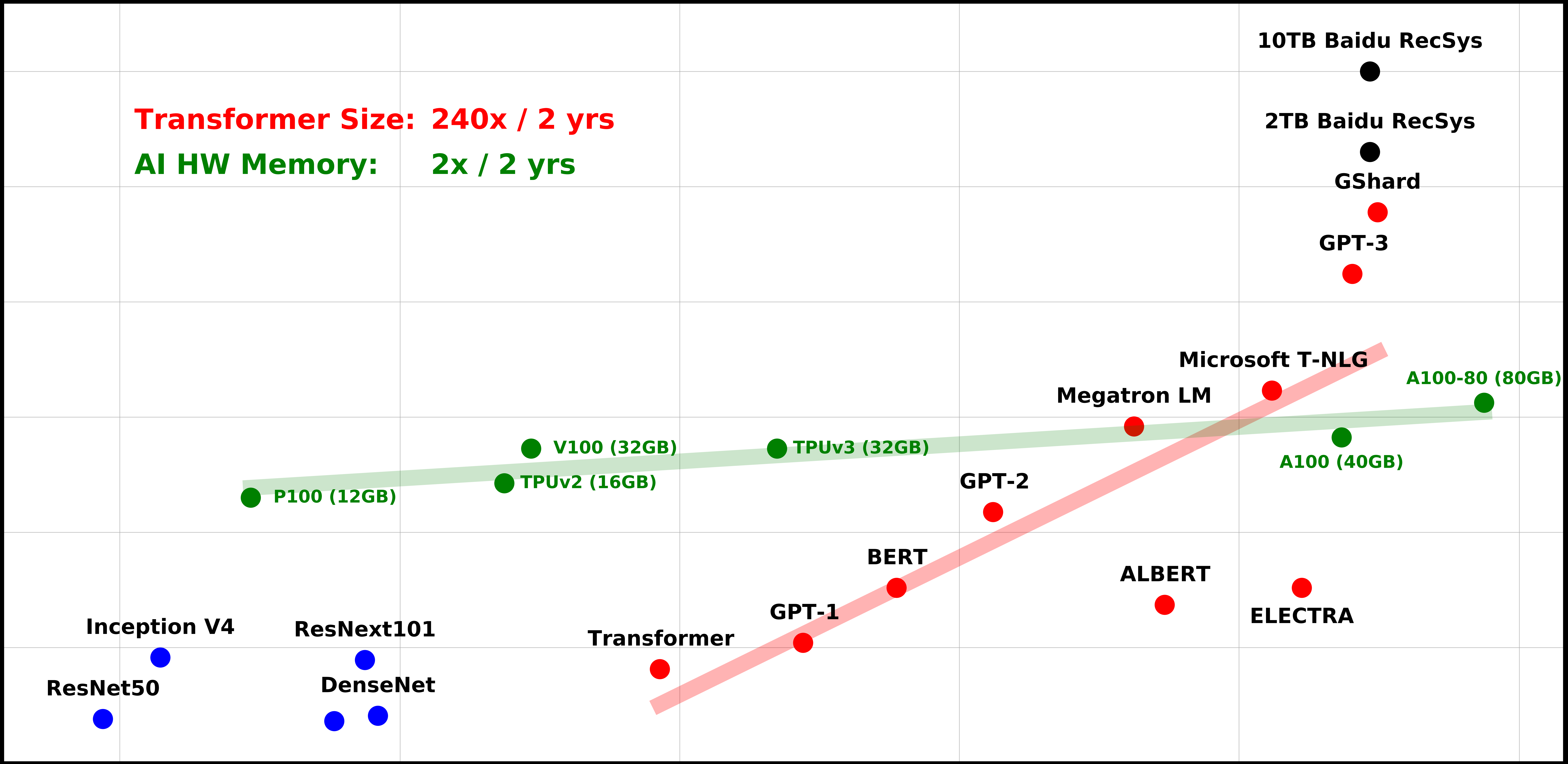
- **2.15^{25} floating point operations**
- **25,000 A100 GPUs**
- **90-100 days**
- **$63 million**
- **Trained on 13 trillion tokens**

semianalysis 2023



Train (GPT-4):
- **2.15^{25} floating point operations**
- **25,000 A100 GPUs**
- **90-100 days**
- **$63 million**
- **Trained on 13 trillion tokens**

Inference (GPT-4):
- **Multiple clusters of 128 GPUs**
- **Model carefully mapped onto hardware**



10TB Baidu RecSys
2TB Baidu RecSys
Transformer Size: 240x / 2 yrs
AI HW Memory: 2x / 2 yrs
GShard
GPT-3
Microsoft T-NLG
A100-80 (80GB)
Megatron LM
V100 (32GB)
TPUv3 (32GB)
A100 (40GB)
TPUv2 (16GB)
GPT-2
P100 (12GB)
BERT
ALBERT
ELECTRA
Inception V4
ResNext101
GPT-1
Transformer
ResNet50
DenseNet

Kaplan et al. (2020)

Resources: *128 interconnected GPUs*
Latency : *10^1 seconds*

You
IEEE style reference please: @ARTICLE{9447722,
author={Guglielmo, Giuseppe Di and Fahim, Farah and Herwig, Christian and Valentin, Manuel Blanco and Duarte, Javier and Gingu, Cristian and Harris, Philip and Hirschauer, James and Kwok, Martin and Loncar, Vladimir and Luo, Yingyi and Miranda, Llovizna and Ngadiuba, Jennifer and Noonan, Daniel and Ogrenci-Memik, Seda and Pierini, Maurizio and Summers, Sioni and Tran, Nhan},
journal={IEEE Transactions on Nuclear Science},
title={A Reconfigurable Neural Network ASIC for Detector Front-End Data Compression at the HL-LHC},
year={2021},
volume={68},
number={8},
pages={2179-2186},
doi={10.1109/TNS.2021.3087100}}

ChatGPT

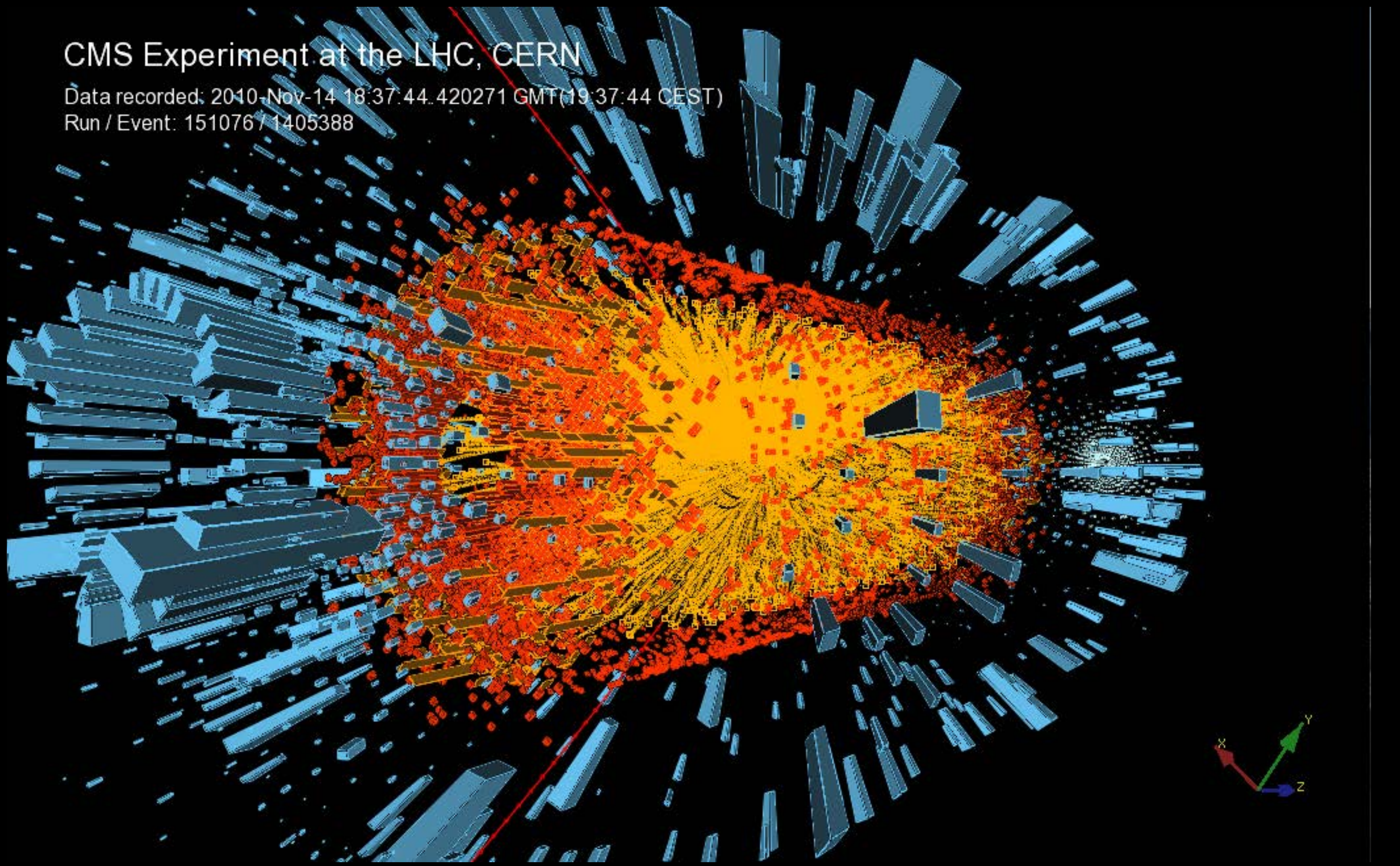
Resources: ***128 interconnected GPUs***
Latency : ***10 seconds***

You

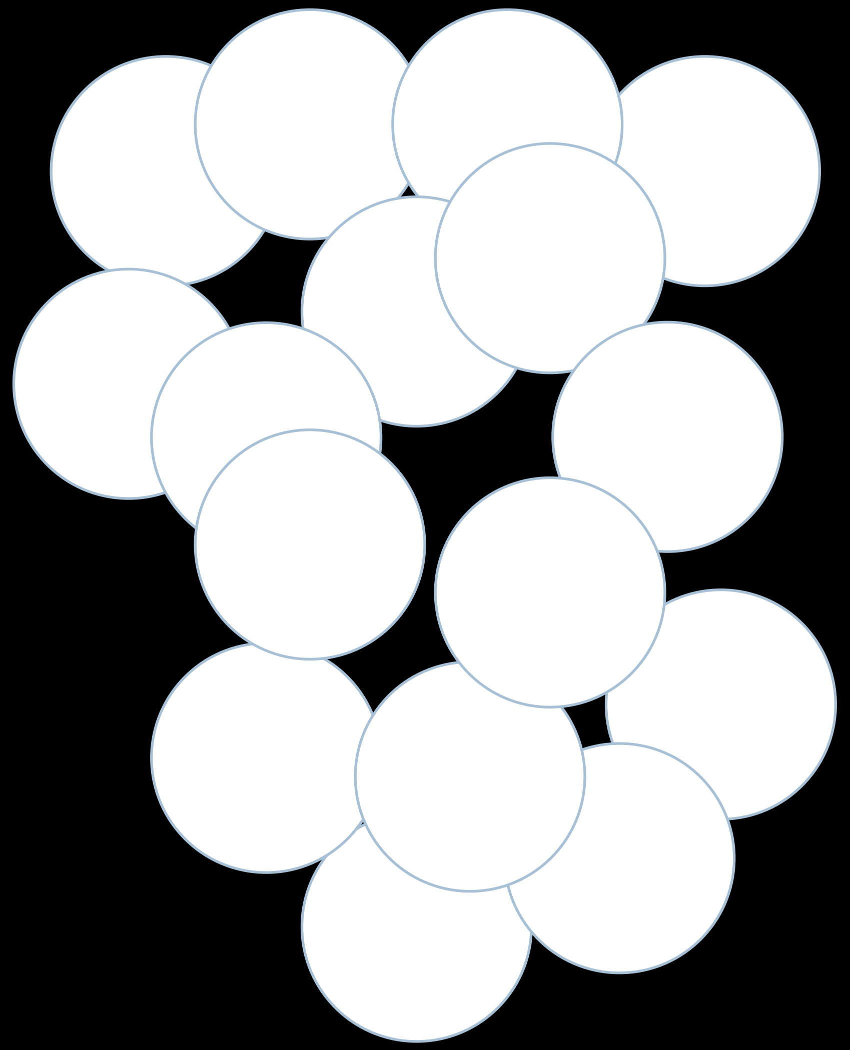
IEEE style reference please: @ARTICLE{9447722,
author={Guglielmo, Giuseppe Di and Fahim, Farah and Herwig, Christian and Valentin, Manuel Blanco and Duarte, Javier and Gingu, Cristian and Harris, Philip and Hirschauer, James and Kwok, Martin and Loncar, Vladimir and Luo, Yingyi and Miranda, Llovizna and Ngadiuba, Jennifer and Noonan, Daniel and Ogrenci-Memik, Seda and Pierini, Maurizio and Summers, Sioni and Tran, Nhan},
journal={IEEE Transactions on Nuclear Science},
title={A Reconfigurable Neural Network ASIC for Detector Front-End Data Compression at the HL-LHC},
year={2021},
volume={68},
number={8},
pages={2179-2186},
doi={10.1109/TNS.2021.3087100}}

ChatGPT

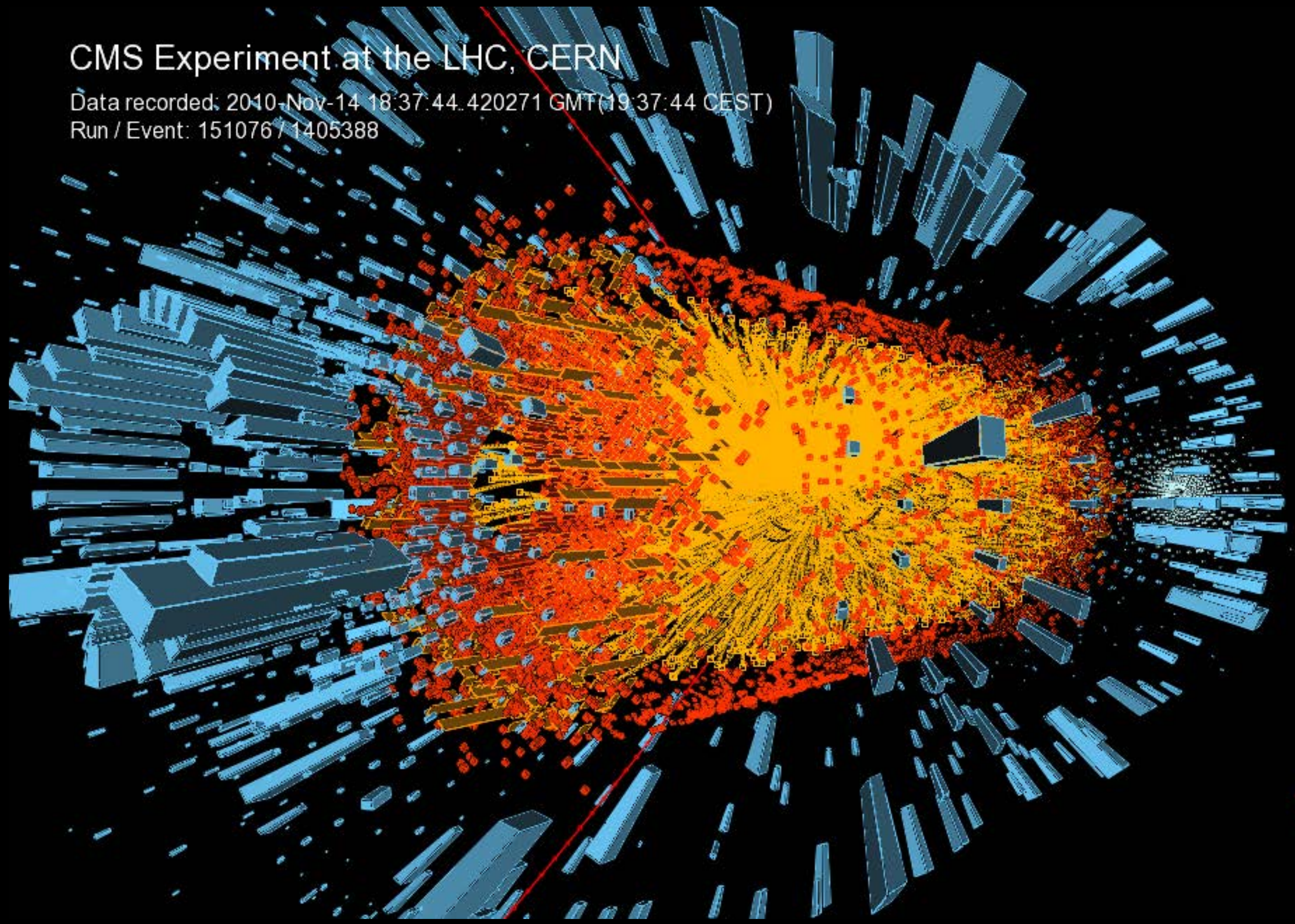
Resources: ***O(10) single chips***
Latency: ***1 millionth of a second***
5% of internet traffic



GPT-4



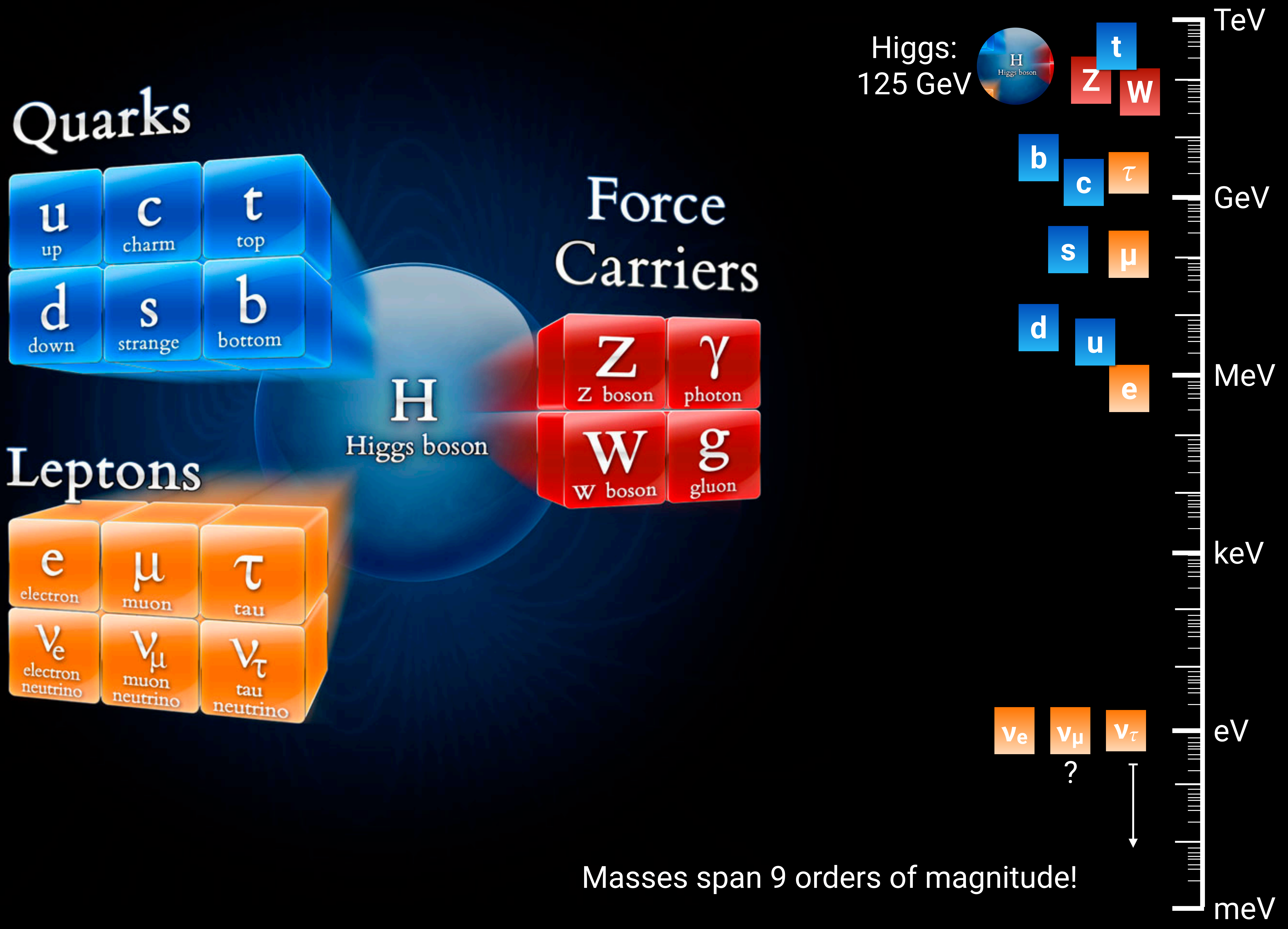
?

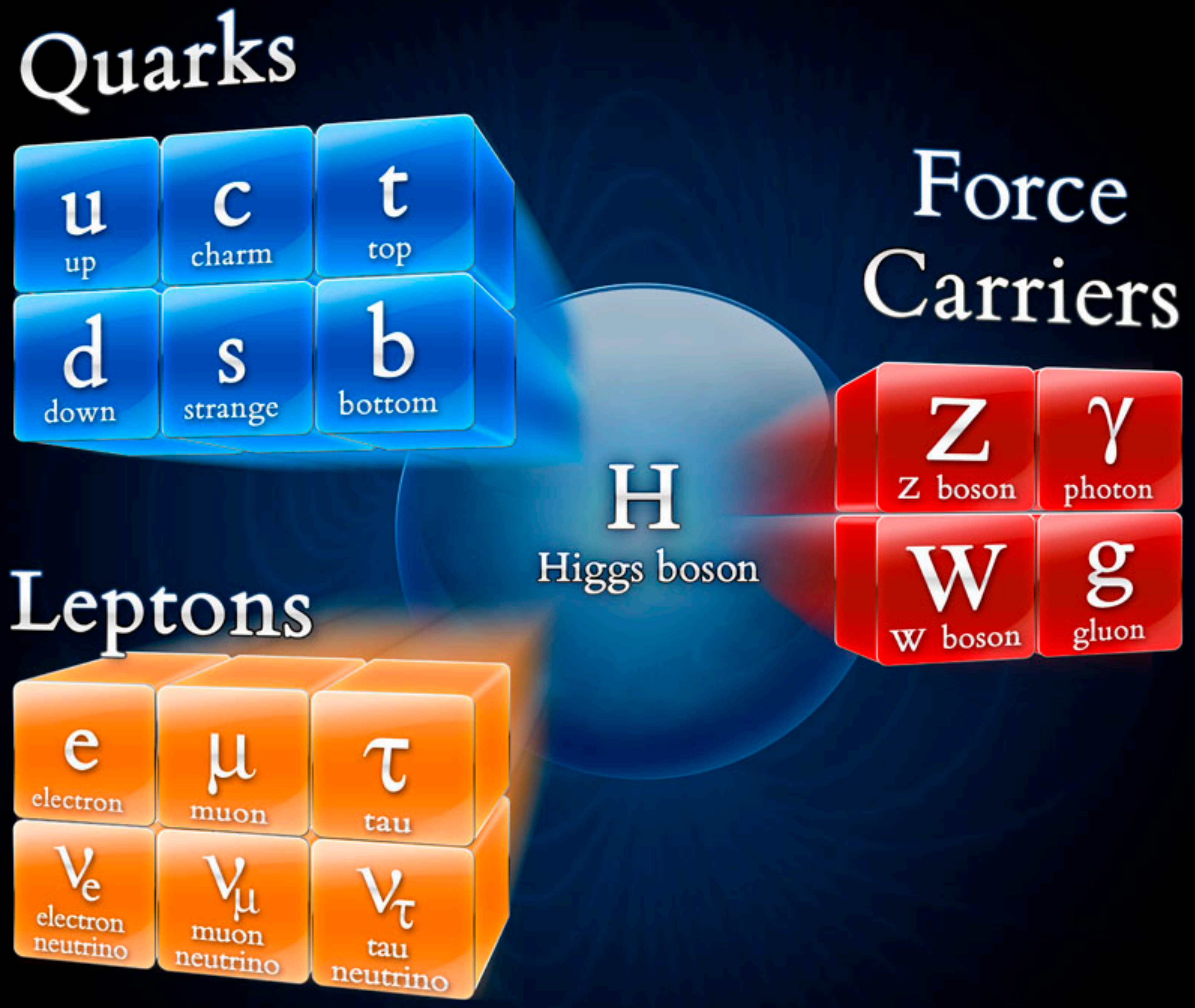

CMS Experiment at the LHC, CERN
Data recorded: 2010-Nov-14 18:37:44.420271 GMT(19:37:44 CEST)
Run / Event: 151076 / 1405388


Geneva
Geneva Lake
LHCb
ATLAS
ALICE
CMS
LHC

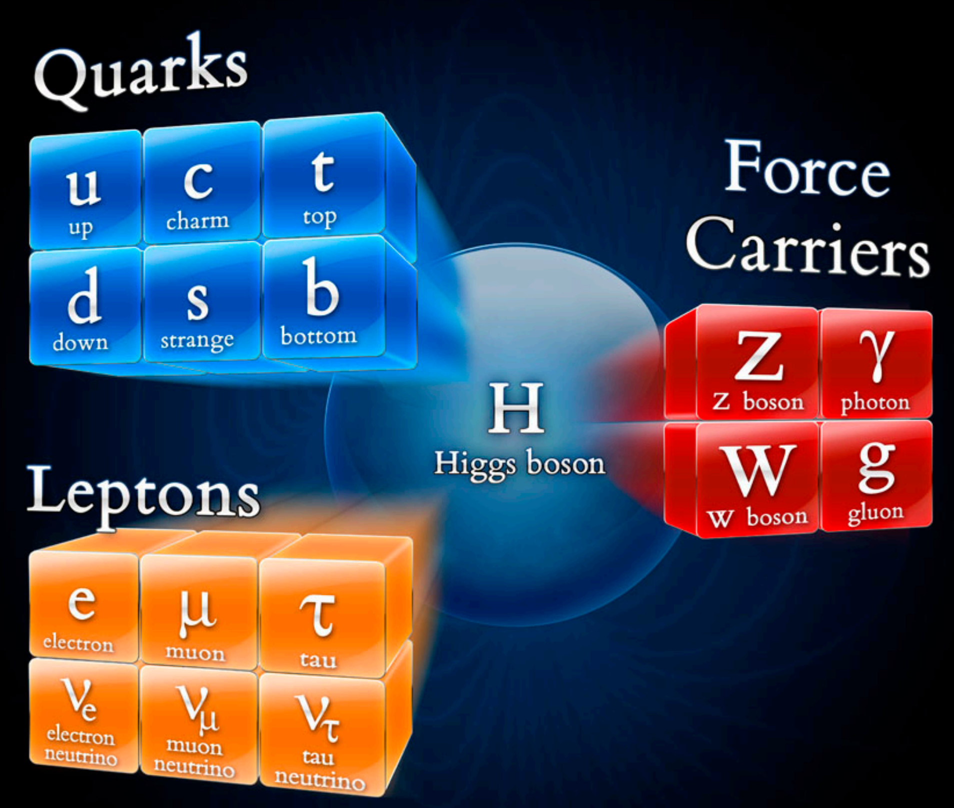

Geneva
Geneva Lake
LHCb
ATLAS
ALICE
CMS
2,500 bunches
10^{11} protons
11,000 times/s
$2.2 \cdot 10^{11}$ protons
25 ns
7.5 m

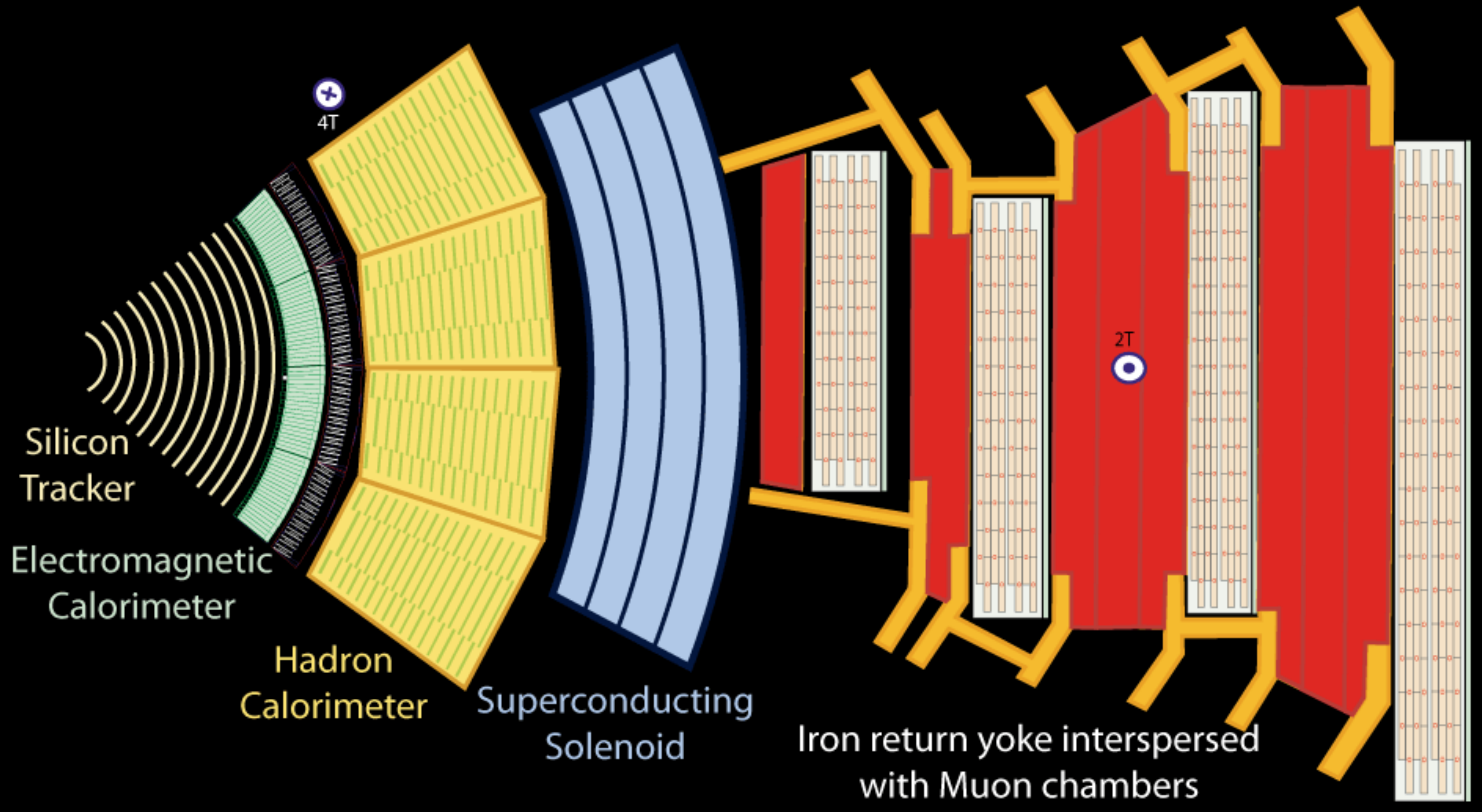

CMS Experiment at the LHC, CERN
Data recorded: 2010-Nov-14 18:37:44.420271 GMT(19:37:44 CEST)
Run / Event: 151076 / 1405388
E = mc²
Y
X
Z

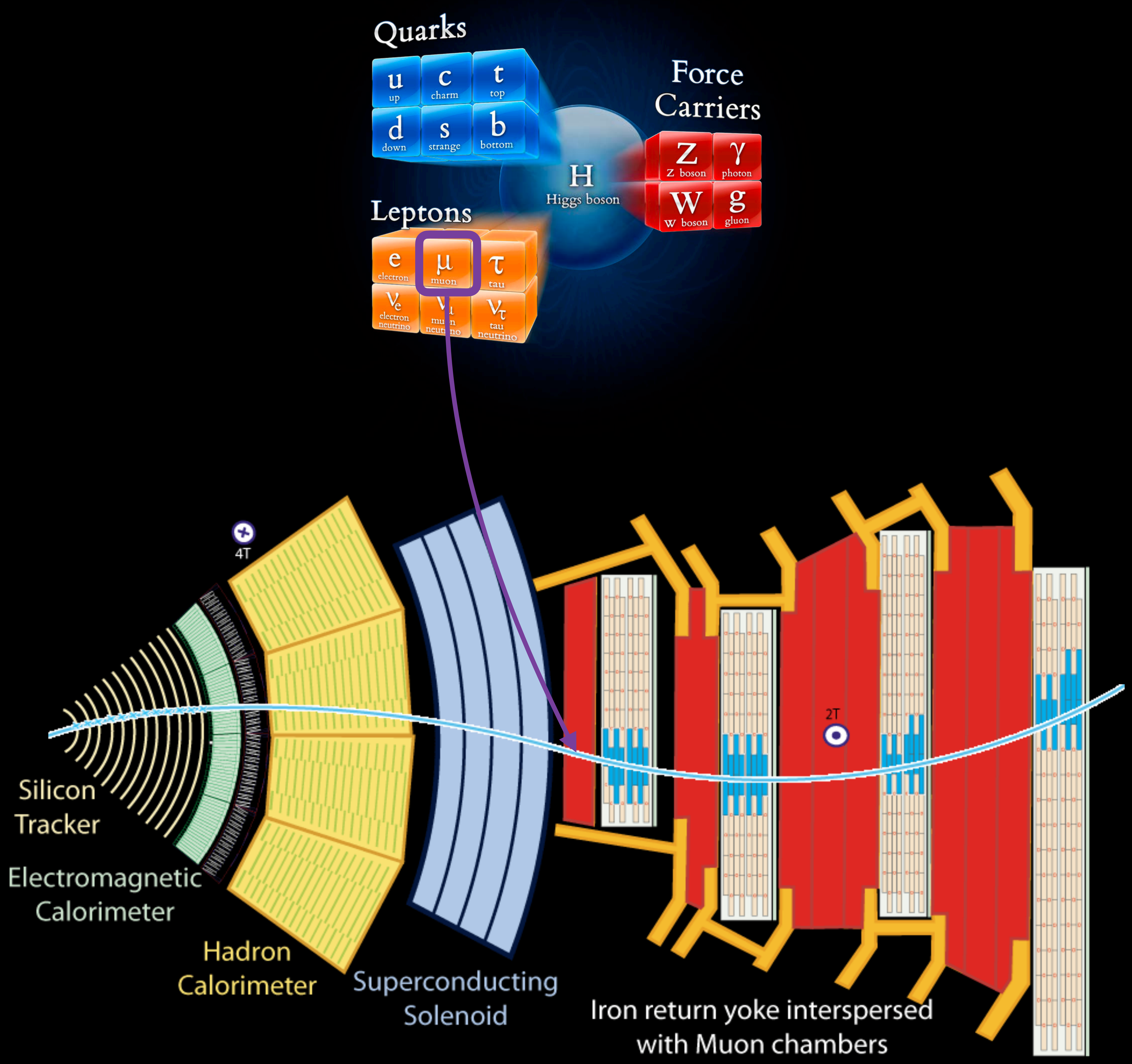

Quarks
u
up
c
charm
t
top
d
down
s
strange
b
bottom
Leptons
e
electron
μ
muon
τ
tau
νe
electron neutrino
νμ
muon neutrino
ντ
tau neutrino
H
Higgs boson
Force Carriers
Z
Z boson
γ
photon
W
W boson
g
gluon
Higgs:
125 GeV
t
Z
W
b
c
τ
s
μ
d
u
e
νe
νμ
ντ
?
TeV
GeV
MeV
keV
eV
meV
Masses span 9 orders of magnitude!

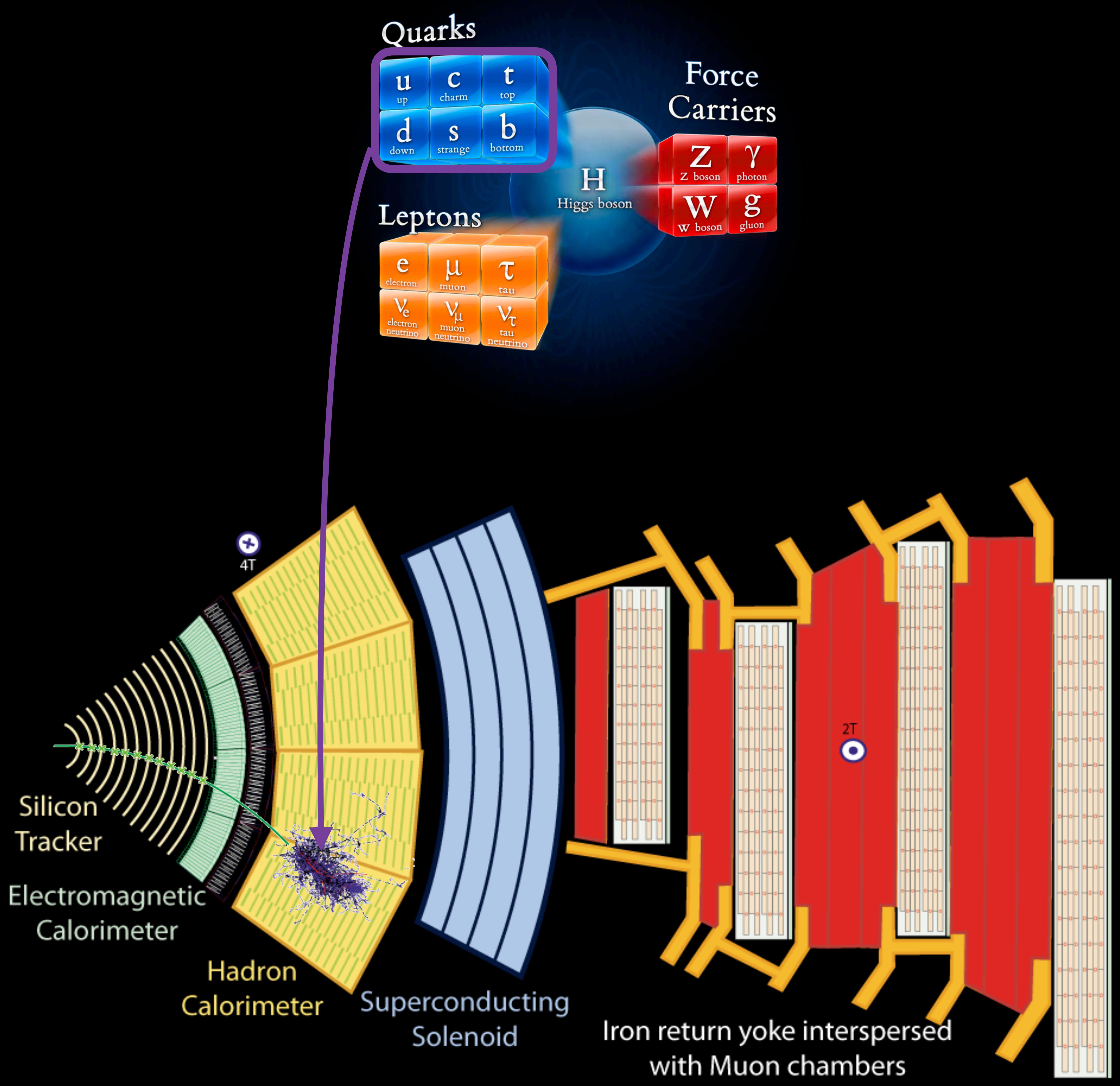


Quarks
u
up
c
charm
t
top
d
down
s
strange
b
bottom
Leptons
e
electron
μ
muon
τ
tau
νe
electron neutrino
νμ
muon neutrino
ντ
tau neutrino
H
Higgs boson
Force Carriers
Z
Z boson
γ
photon
W
W boson
g
gluon


Quarks
u
up
c
charm
t
top
d
down
s
strange
b
bottom
Leptons
e
electron
μ
muon
τ
tau
νe
electron neutrino
νμ
muon neutrino
ντ
tau neutrino
H
Higgs boson
Force Carriers
Z
Z boson
γ
photon
W
W boson
g
gluon


4T
2T
Silicon Tracker
Electromagnetic Calorimeter
Hadron Calorimeter
Superconducting Solenoid
Iron return yoke interspersed with Muon chambers

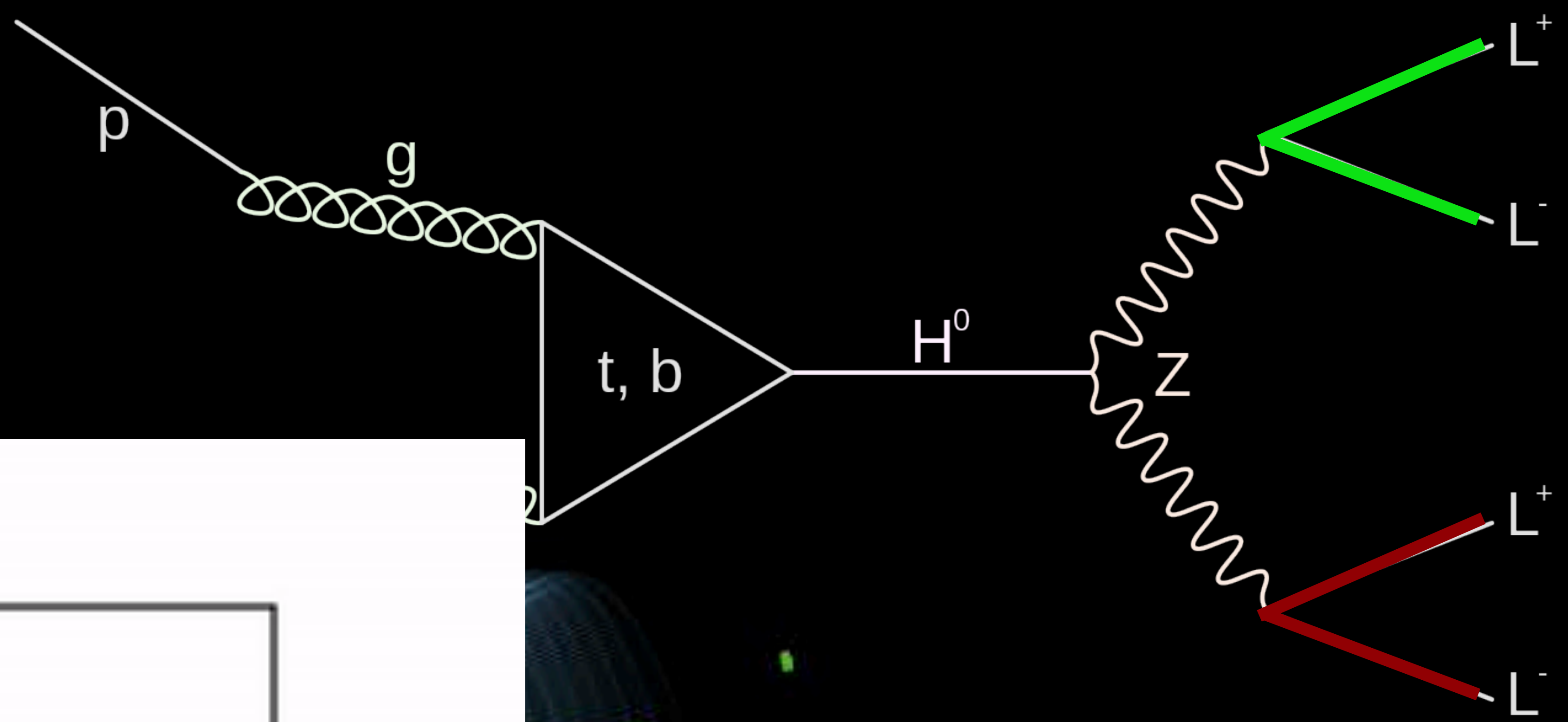

Quarks
u up
c charm
t top
d down
s strange
b bottom
Force Carriers
Z Z boson
γ photon
W W boson
g gluon
H Higgs boson
Leptons
e electron
μ muon
τ tau
νe electron neutrino
ντ tau neutrino
4T
2T
Silicon Tracker
Electromagnetic Calorimeter
Hadron Calorimeter
Superconducting Solenoid
Iron return yoke interspersed with Muon chambers

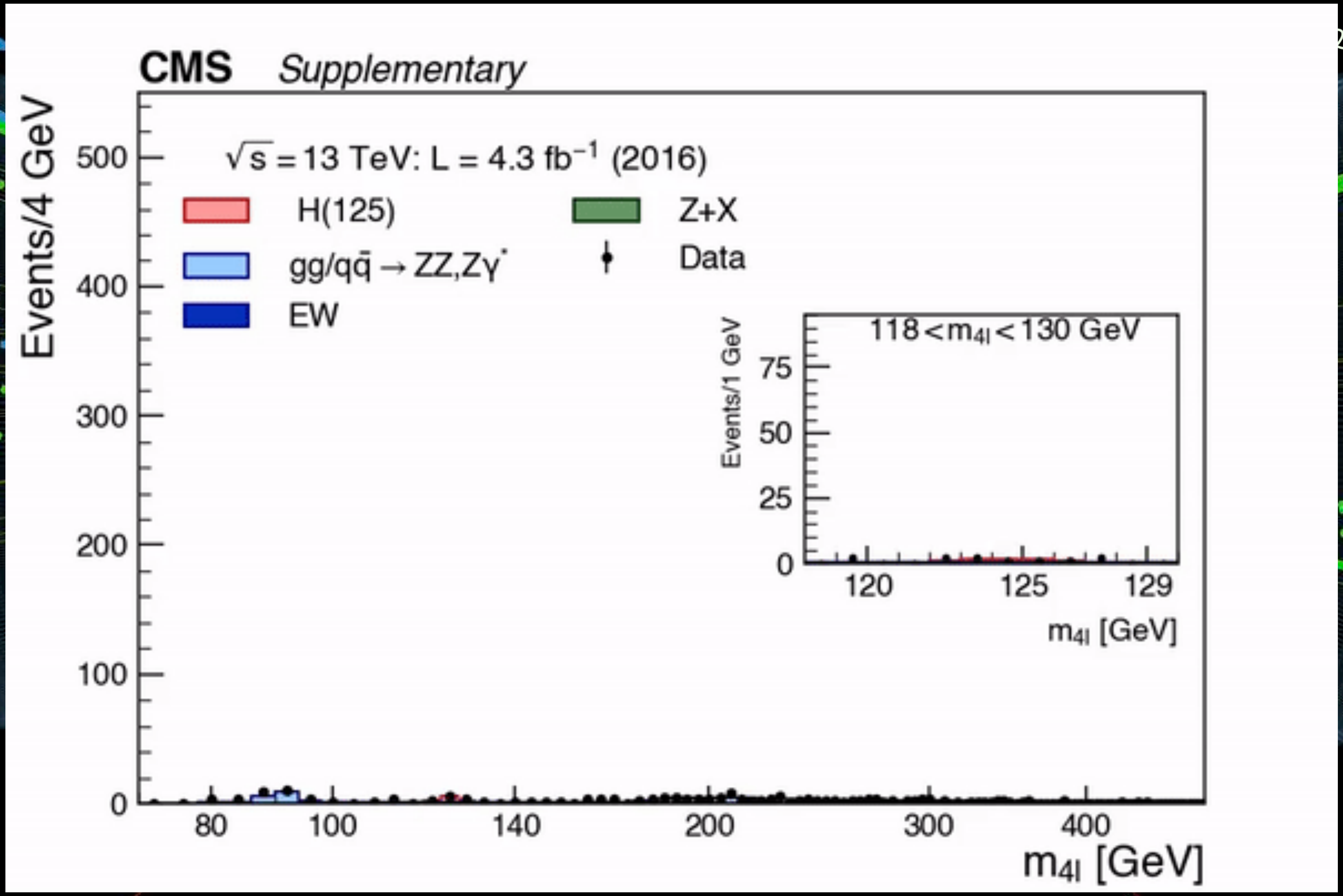


Quarks
u up
c charm
t top
d down
s strange
b bottom
Leptons
e electron
μ muon
τ tau
νe electron neutrino
νμ muon neutrino
ντ tau neutrino
H Higgs boson
Force Carriers
Z Z boson
γ photon
W W boson
g gluon
4T
2T
Silicon Tracker
Electromagnetic Calorimeter
Hadron Calorimeter
Superconducting Solenoid
Iron return yoke interspersed with Muon chambers


p
g
t, b
H^0
Z
L^+
L^-
L^+
L^-
g
p

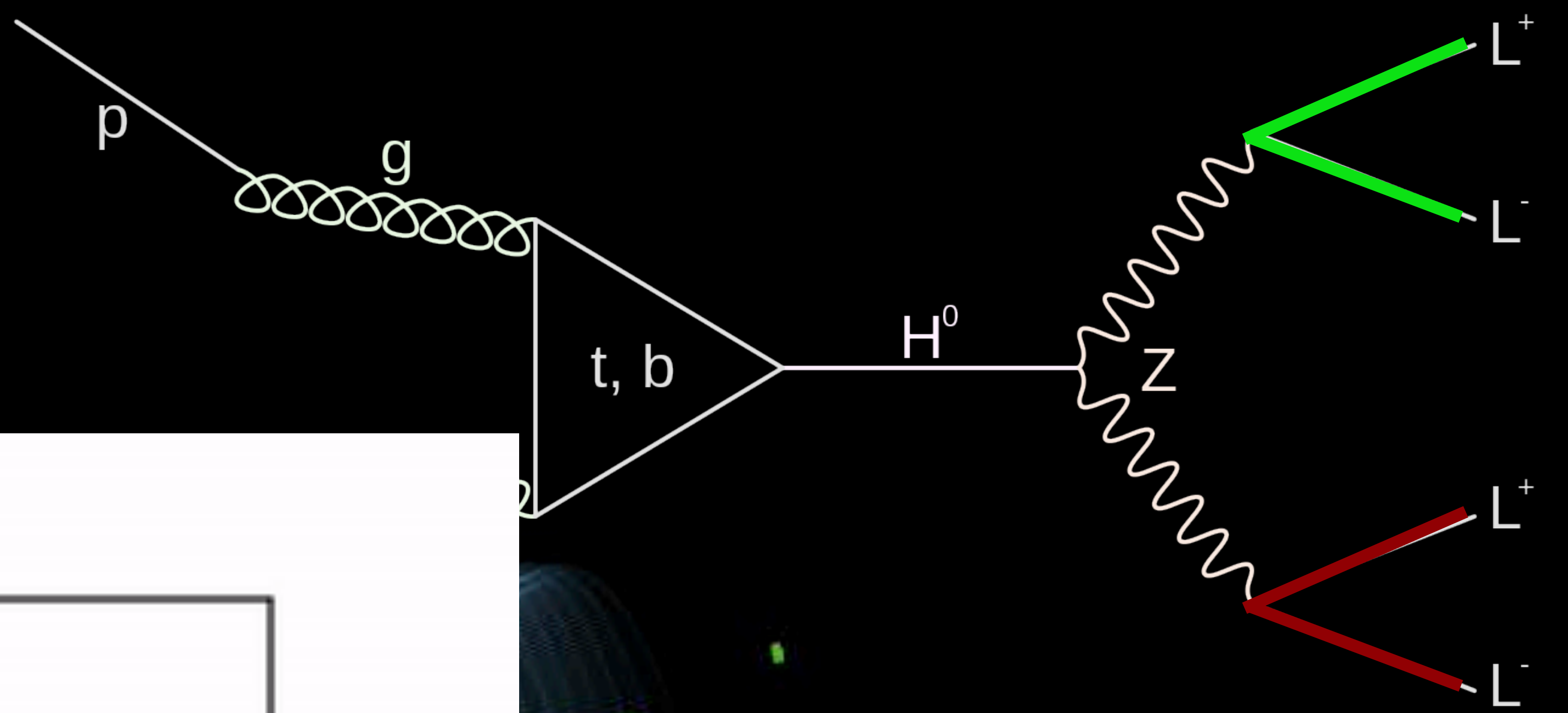
cmsexperiment.web.cern.ch

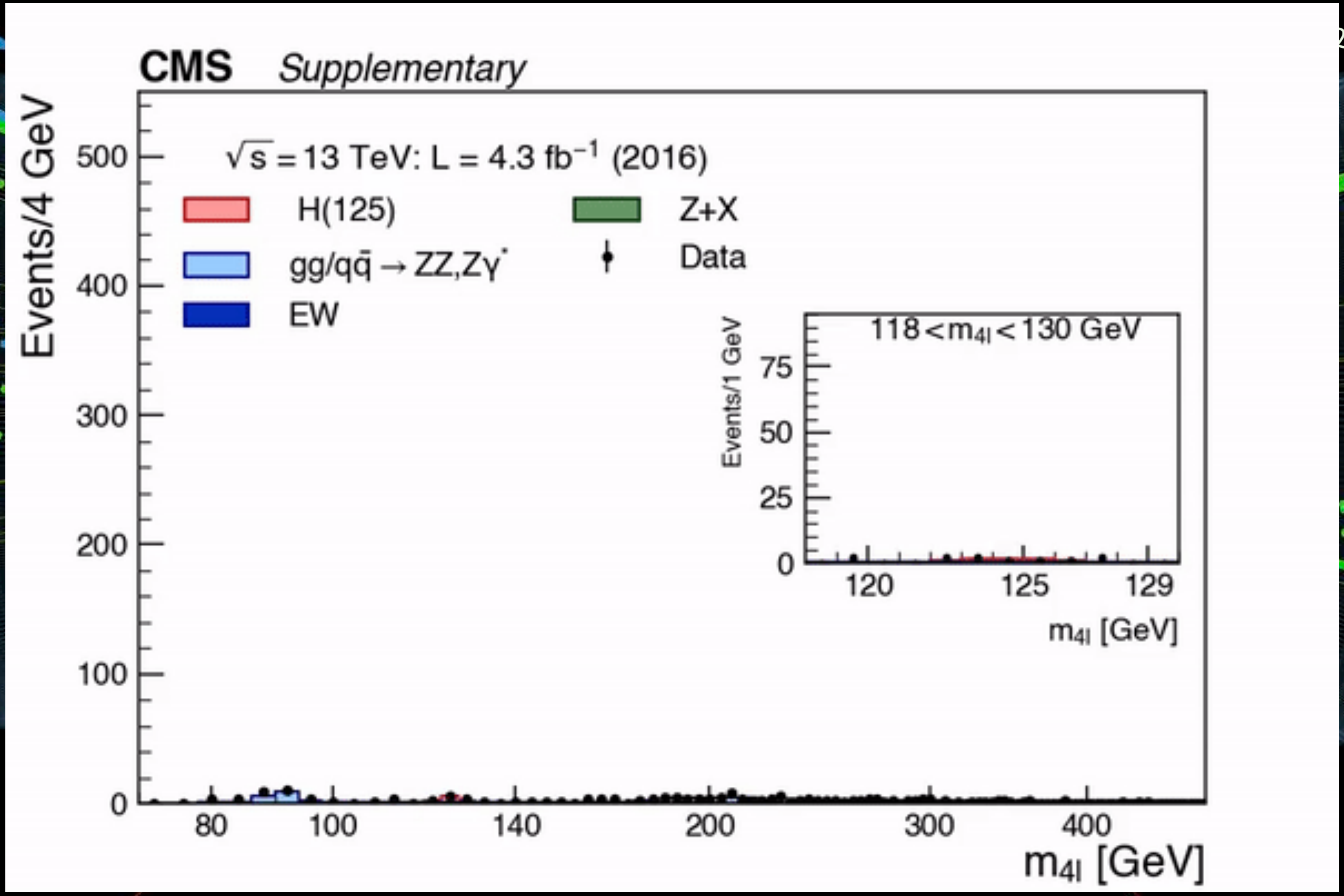




We had to collide billions of protons,
only around 10 signal events were needed to claim discovery!

cmsexperiment.web.cern.ch





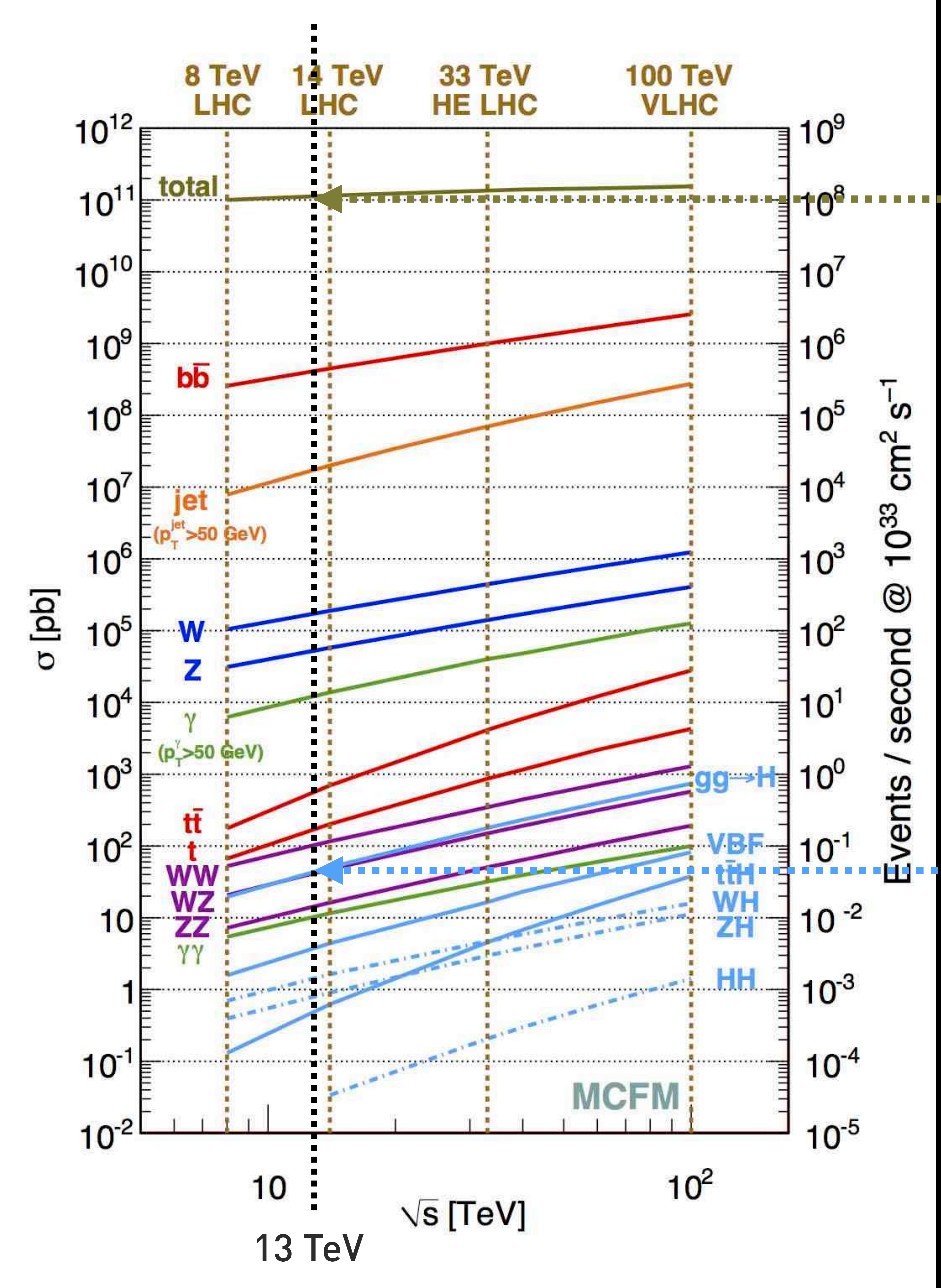
We had to collide billions of protons,
only around 10 signal events were needed to claim discovery!

The Standard Model

$$
\begin{aligned}
&-\tfrac12\partial_\nu g^a_\mu\partial_\nu g^a_\mu - g_s f^{abc}\partial_\mu g^a_\nu g^b_\mu g^c_\nu - \tfrac14 g_s^2 f^{abc}f^{ade}g^b_\mu g^c_\nu g^d_\mu g^e_\nu + \\
&\tfrac12 ig_s^2(\bar q^\sigma_i\gamma^\mu q^\sigma_j)g^a_\mu + \bar G^a\partial^2 G^a + g_s f^{abc}\partial_\mu\bar G^a G^b g^c_\mu - \partial_\nu W^+_\mu\partial_\nu W^-_\mu - \\
&M^2W^+_\mu W^-_\mu - \tfrac12\partial_\nu Z^0_\mu\partial_\nu Z^0_\mu - \tfrac{1}{2c_w^2}M^2Z^0_\mu Z^0_\mu - \tfrac12\partial_\mu A_\nu\partial_\mu A_\nu - \tfrac12\partial_\mu H\partial_\mu H - \\
&\tfrac12 m_h^2H^2 - \partial_\mu\phi^+\partial_\mu\phi^- - M^2\phi^+\phi^- - \tfrac12\partial_\mu\phi^0\partial_\mu\phi^0 - \tfrac{1}{2c_w^2}M\phi^0\phi^0 - \beta_h[\tfrac{2M^2}{g^2} + \\
&\tfrac{2M}{g}H + \tfrac12(H^2+\phi^0\phi^0+2\phi^+\phi^-)] + \tfrac{2M^4}{g^2}\alpha_h - igc_w[\partial_\nu Z^0_\mu(W^+_\mu W^-_\nu - \\
&W^+_\nu W^-_\mu) - Z^0_\nu(W^+_\mu\partial_\nu W^-_\mu - W^-_\mu\partial_\nu W^+_\mu) + Z^0_\mu(W^+_\nu\partial_\nu W^-_\mu - \\
&W^-_\nu\partial_\nu W^+_\mu)] - igs_w[\partial_\nu A_\mu(W^+_\mu W^-_\nu - W^+_\nu W^-_\mu) - A_\nu(W^+_\mu\partial_\nu W^-_\mu - \\
&W^-_\mu\partial_\nu W^+_\mu) + A_\mu(W^+_\nu\partial_\nu W^-_\mu - W^-_\nu\partial_\nu W^+_\mu)] - \tfrac12 g^2W^+_\mu W^-_\mu W^+_\nu W^-_\nu + \\
&\tfrac12 g^2W^+_\mu W^-_\nu W^+_\mu W^-_\nu + g^2c_w^2(Z^0_\mu W^+_\mu Z^0_\nu W^-_\nu - Z^0_\mu Z^0_\mu W^+_\nu W^-_\nu) + \\
&g^2s_w^2(A_\mu W^+_\mu A_\nu W^-_\nu - A_\mu A_\mu W^+_\nu W^-_\nu) + g^2s_wc_w[A_\mu Z^0_\nu(W^+_\mu W^-_\nu - \\
&W^+_\nu W^-_\mu) - 2A_\mu Z^0_\mu W^+_\nu W^-_\nu] - g\alpha[H^3 + H\phi^0\phi^0 + 2H\phi^+\phi^-] - \\
&\tfrac18 g^2\alpha_h[H^4+(\phi^0)^4+4(\phi^+\phi^-)^2+4(\phi^0)^2\phi^+\phi^-+4H^2\phi^+\phi^-+2(\phi^0)^2H^2] - \\
&gMW^+_\mu W^-_\mu H - \tfrac12 g\tfrac{M}{c_w^2}Z^0_\mu Z^0_\mu H - \tfrac12 ig[W^+_\mu(\phi^0\partial_\mu\phi^- - \phi^-\partial_\mu\phi^0) - \\
&W^-_\mu(\phi^0\partial_\mu\phi^+ - \phi^+\partial_\mu\phi^0)] + \tfrac12 g[W^+_\mu(H\partial_\mu\phi^- - \phi^-\partial_\mu H) - W^-_\mu(H\partial_\mu\phi^+ - \\
&\phi^+\partial_\mu H)] + \tfrac12 g\tfrac{1}{c_w}(Z^0_\mu(H\partial_\mu\phi^0 - \phi^0\partial_\mu H) - ig\tfrac{s_w^2}{c_w}MZ^0_\mu(W^+_\mu\phi^- - W^-_\mu\phi^+) + \\
&igs_wMA_\mu(W^+_\mu\phi^- - W^-_\mu\phi^+) - ig\tfrac{1-2c_w^2}{2c_w}Z^0_\mu(\phi^+\partial_\mu\phi^- - \phi^-\partial_\mu\phi^+) + \\
&igs_wA_\mu(\phi^+\partial_\mu\phi^- - \phi^-\partial_\mu\phi^+) - \tfrac14 g^2W^+_\mu W^-_\mu[H^2+(\phi^0)^2+2\phi^+\phi^-] - \\
&\tfrac14 g^2\tfrac{1}{c_w^2}Z^0_\mu Z^0_\mu[H^2+(\phi^0)^2+2(2s_w^2-1)^2\phi^+\phi^-] - \tfrac12 g^2\tfrac{s_w^2}{c_w}Z^0_\mu\phi^0(W^+_\mu\phi^- + \\
&W^-_\mu\phi^+) - \tfrac12 ig^2\tfrac{s_w^2}{c_w}Z^0_\mu H(W^+_\mu\phi^- - W^-_\mu\phi^+) + \tfrac12 g^2s_wA_\mu\phi^0(W^+_\mu\phi^- + \\
&W^-_\mu\phi^+) + \tfrac12 ig^2s_wA_\mu H(W^+_\mu\phi^- - W^-_\mu\phi^+) - g^2\tfrac{s_w}{c_w}(2c_w^2-1)Z^0_\mu A_\mu\phi^+\phi^- - \\
&g^1s_w^2A_\mu A_\mu\phi^+\phi^- - \bar e^\lambda(\gamma\partial+m_e^\lambda)e^\lambda - \bar\nu^\lambda\gamma\partial\nu^\lambda - \bar u^\lambda_j(\gamma\partial+m_u^\lambda)u^\lambda_j - \\
&\bar d^\lambda_j(\gamma\partial+m_d^\lambda)d^\lambda_j + igs_wA_\mu[-(\bar e^\lambda\gamma^\mu e^\lambda) + \tfrac23(\bar u^\lambda_j\gamma^\mu u^\lambda_j) - \tfrac13(\bar d^\lambda_j\gamma^\mu d^\lambda_j)] + \\
&\tfrac{ig}{4c_w}Z^0_\mu[(\bar\nu^\lambda\gamma^\mu(1+\gamma^5)\nu^\lambda) + (\bar e^\lambda\gamma^\mu(4s_w^2-1-\gamma^5)e^\lambda) + (\bar u^\lambda_j\gamma^\mu(\tfrac43 s_w^2 - \\
&1-\gamma^5)u^\lambda_j) + (\bar d^\lambda_j\gamma^\mu(1-\tfrac83 s_w^2-\gamma^5)d^\lambda_j)] + \tfrac{ig}{2\sqrt2}W^+_\mu[(\bar\nu^\lambda\gamma^\mu(1+\gamma^5)e^\lambda) + \\
&(\bar u^\lambda_j\gamma^\mu(1+\gamma^5)C_{\lambda\kappa}d^\kappa_j)] + \tfrac{ig}{2\sqrt2}W^-_\mu[(\bar e^\lambda\gamma^\mu(1+\gamma^5)\nu^\lambda) + (\bar d^\kappa_j C^\dagger_{\lambda\kappa}\gamma^\mu(1+ \\
&\gamma^5)u^\lambda_j)] + \tfrac{ig}{2\sqrt2}\tfrac{m_e^\lambda}{M}[-\phi^+(\bar\nu^\lambda(1-\gamma^5)e^\lambda) + \phi^-(\bar e^\lambda(1+\gamma^5)\nu^\lambda)] - \\
&\tfrac g2\tfrac{m_e^\lambda}{M}[H(\bar e^\lambda e^\lambda) + i\phi^0(\bar e^\lambda\gamma^5 e^\lambda)] + \tfrac{ig}{2M\sqrt2}\phi^+[-m_d^\kappa(\bar u^\lambda_j C_{\lambda\kappa}(1-\gamma^5)d^\kappa_j) + \\
&m_u^\lambda(\bar u^\lambda_j C_{\lambda\kappa}(1+\gamma^5)d^\kappa_j] + \tfrac{ig}{2M\sqrt2}\phi^-[m_d^\lambda(\bar d^\lambda_j C^\dagger_{\lambda\kappa}(1+\gamma^5)u^\kappa_j) - m_u^\kappa(\bar d^\lambda_j C^\dagger_{\lambda\kappa}(1- \\
&\gamma^5)u^\kappa_j] - \tfrac g2\tfrac{m_u^\lambda}{M}H(\bar u^\lambda_j u^\lambda_j) - \tfrac g2\tfrac{m_d^\lambda}{M}H(\bar d^\lambda_j d^\lambda_j) + \tfrac{ig}{2}\tfrac{m_u^\lambda}{M}\phi^0(\bar u^\lambda_j\gamma^5 u^\lambda_j) - \\
&\tfrac{ig}{2}\tfrac{m_d^\lambda}{M}\phi^0(\bar d^\lambda_j\gamma^5 d^\lambda_j) + \bar X^+(\partial^2-M^2)X^+ + \bar X^-(\partial^2-M^2)X^- + \bar X^0(\partial^2 - \\
&\tfrac{M^2}{c_w^2})X^0 + \bar Y\partial^2 Y + igc_wW^+_\mu(\partial_\mu\bar X^0X^- - \partial_\mu\bar X^+X^0) + igs_wW^+_\mu(\partial_\mu\bar YX^- - \\
&\partial_\mu\bar X^+Y) + igc_wW^-_\mu(\partial_\mu\bar X^-X^0 - \partial_\mu\bar X^0X^+) + igs_wW^-_\mu(\partial_\mu\bar X^-Y - \\
&\partial_\mu\bar YX^+) + igc_wZ^0_\mu(\partial_\mu\bar X^+X^+ - \partial_\mu\bar X^-X^-) + igs_wA_\mu(\partial_\mu\bar X^+X^+ - \\
&\partial_\mu\bar X^-X^-) - \tfrac12 gM[\bar X^+X^+H + \bar X^-X^-H + \tfrac{1}{c_w^2}\bar X^0X^0H] + \\
&\tfrac{1-2c_w^2}{2c_w}igM[\bar X^+X^0\phi^+ - \bar X^-X^0\phi^-] + \tfrac{1}{2c_w}igM[\bar X^0X^-\phi^+ - \bar X^0X^+\phi^-] + \\
&igMs_w[\bar X^0X^-\phi^+ - \bar X^0X^+\phi^-] + \tfrac12 igM[\bar X^+X^+\phi^0 - \bar X^-X^-\phi^0]
\end{aligned}
$$

Geneva
Geneva Lake
LHCb
ATLAS
ALICE
CMS
O(1) billion collisions per second
O(1) PB of data per second

"Probability" of producing "anything"



"Probability" of producing a Higgs

Higgs produced
~1 in a billion collisions!

Saving all collisions not useful
(even if we could)!


Geneva
Geneva Lake
LHCb
ATLAS
ALICE
CMS
2 step rate reduction
(hardware+software)
LHC
2 step rate reduction
(hardware+software)


Geneva
Geneva Lake
LHCb
ATLAS
ALICE
CMS
Software rate reduction
(GPU+CPU)
2 step rate reduction
(hardware+software)
Continous read-out
(CPU+GPU)
2 step rate reduction
(hardware+software)
LHC

Geneva
Geneva Lake
LHCb
ATLAS
ALICE
CMS
2 step rate reduction
(hardware+software)
LHC

Geneva
Geneva Lake
LHCb
CMS
Data temporarily stored
in detector electronics for 4 μs
(frontend buffering limit)
LHC


Geneva
Geneva Lake
LHCb
AT
CMS
LHC
5% of internet traffic to L1
Data temporarily stored
in detector electronics for 4 μs
(frontend buffering limit)
L1 trigger:
~1000 FPGAs
Decide which
event to keep
within ~4 μs
Reject >99% of
collisions!

Geneva
Geneva Lake
LHCb
CMS
LHC
L1 bit:
Accept = 1
Reject = 0


Geneva
Geneva Lake
LHCb
ATLAS
ALICE
CMS
L1 accept:
O(100) kHz
~Tb/s
LHC

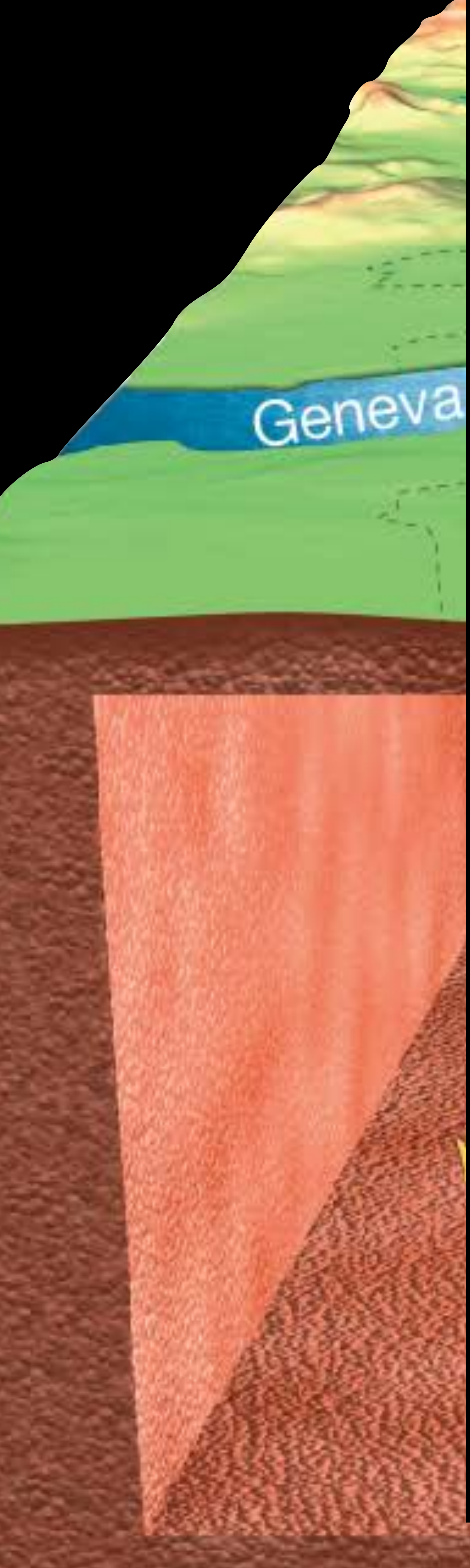

Geneva
Geneva Lake
LHCb
ATLAS
ALICE
CMS
L1 accept:
O(100) kHz
~Tb/s
LHC

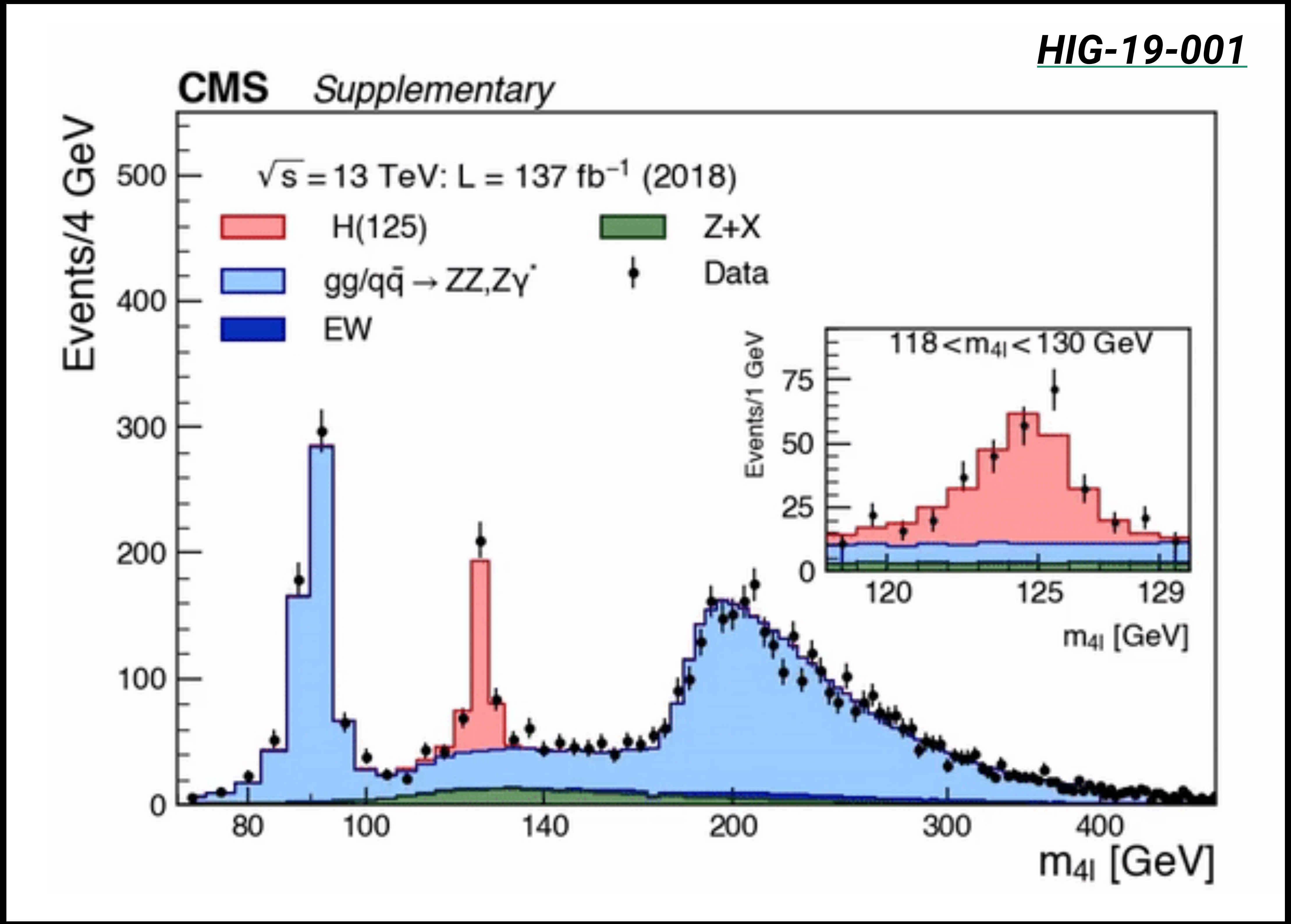

High Level Trigger:
25’600 CPUs / 400 GPUs
Latency: 3-400 ms
Reject further 99%!
eva
LHCb
ATLAS
ALICE
L1 accept:
O(100) kHz
~Tb/s
LHC

TIER 0: ∞
Geneva
High Level Trigger:
Latency O(100) ms
LHCb
ATLAS
HLT accept:
O(1) kHz
~Gb/s
CMS
LHC


TIER 0: ∞
Geneva
High Level Trigger:
Latency O(100) ms
LHCb
ATLAS
HLT accept:
O(1) kHz
~Gb/s
CMS
LHC


TIER 0: ∞
Geneva
High Level Trigger:
Latency O(100) ms
LHCb
ATLAS
0.0025% of collision events remaining
HLT accept:
O(1) kHz
~Gb/s
CMS
LHC

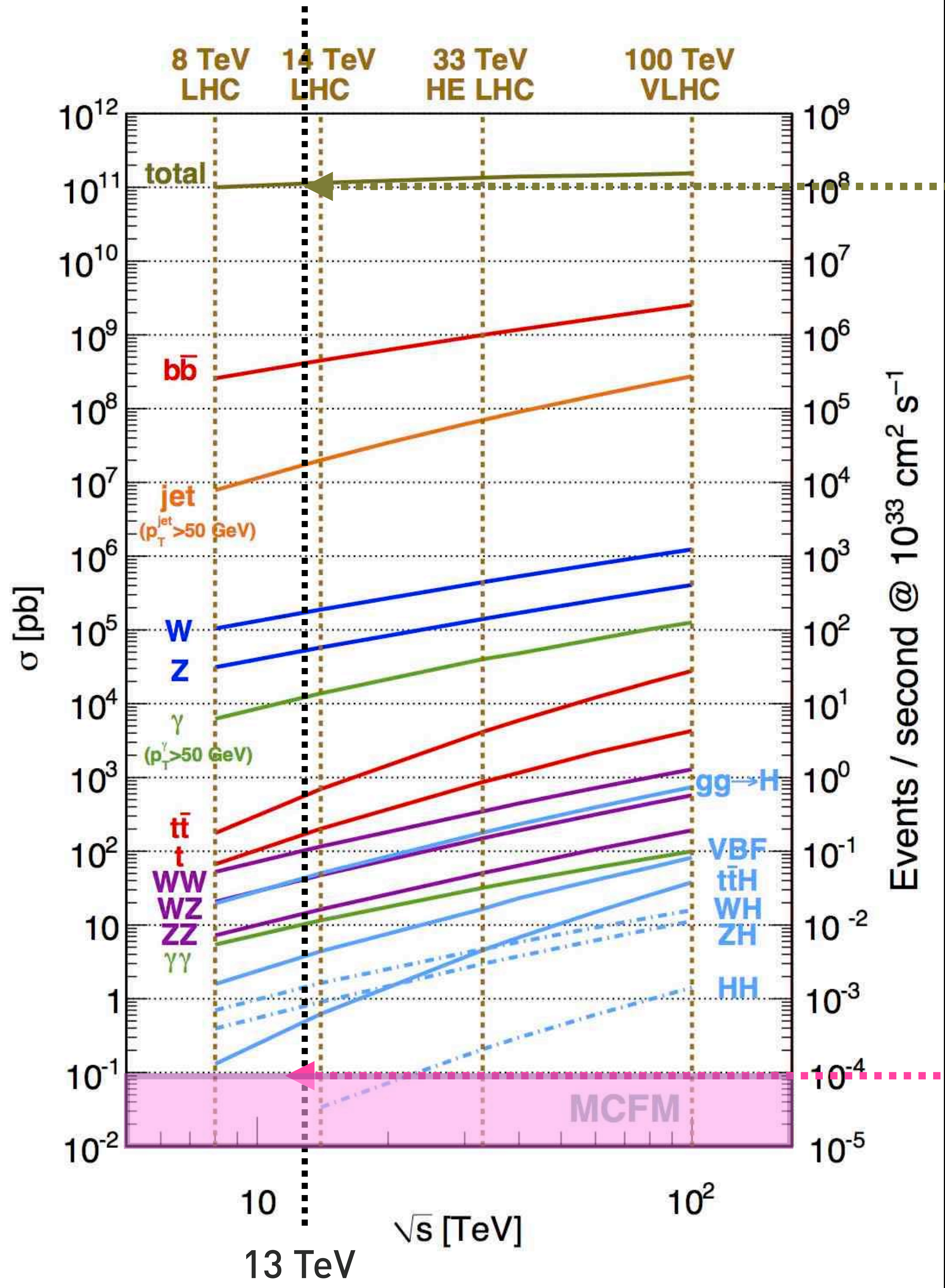




To make sure we select "the right" 0.0025%, algorithms must be

- Fast (get more data through)
- Accurate (select the right data)

CMS Preliminary
LHCP 2017
Vector mediator
EXCLUDED
(a) CMS
DM Simplified Model Exclusions
ATLAS Preliminary March 2017
(b) ATLAS
2016 diboson combination
CMS
Combination
LHCP 2020
Axial-vector mediator
Fall 2020
Z'→qq̄



"Probability" of producing "anything"

New Physics?

New Physics is produced less than
1 in a trillion (if at all)

Need <u>more</u> data!

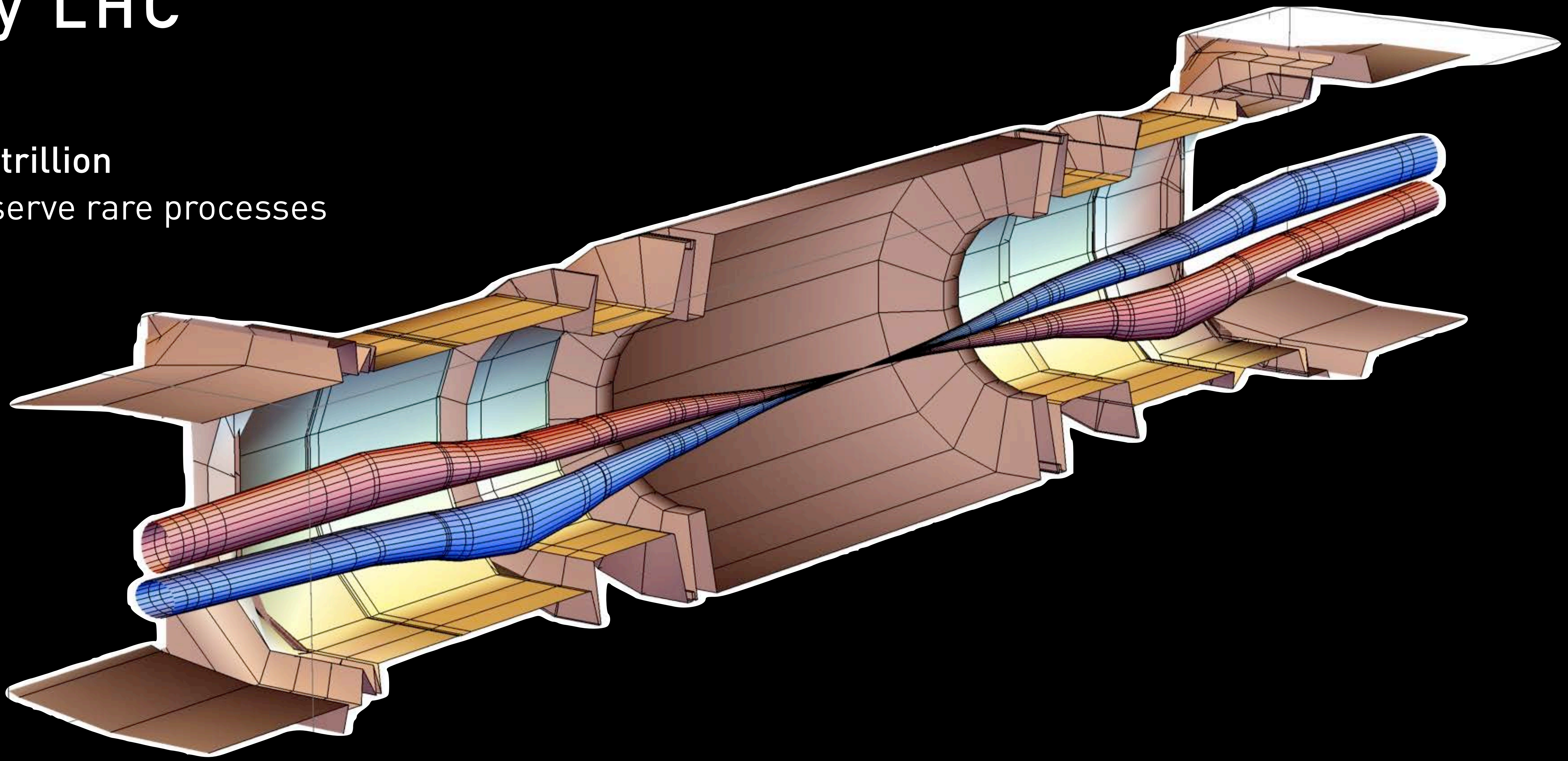
High Luminosity LHC

New Physics is produced 1 in a trillion

- Need <u>more collisions</u> to observe rare processes

High Luminosity LHC

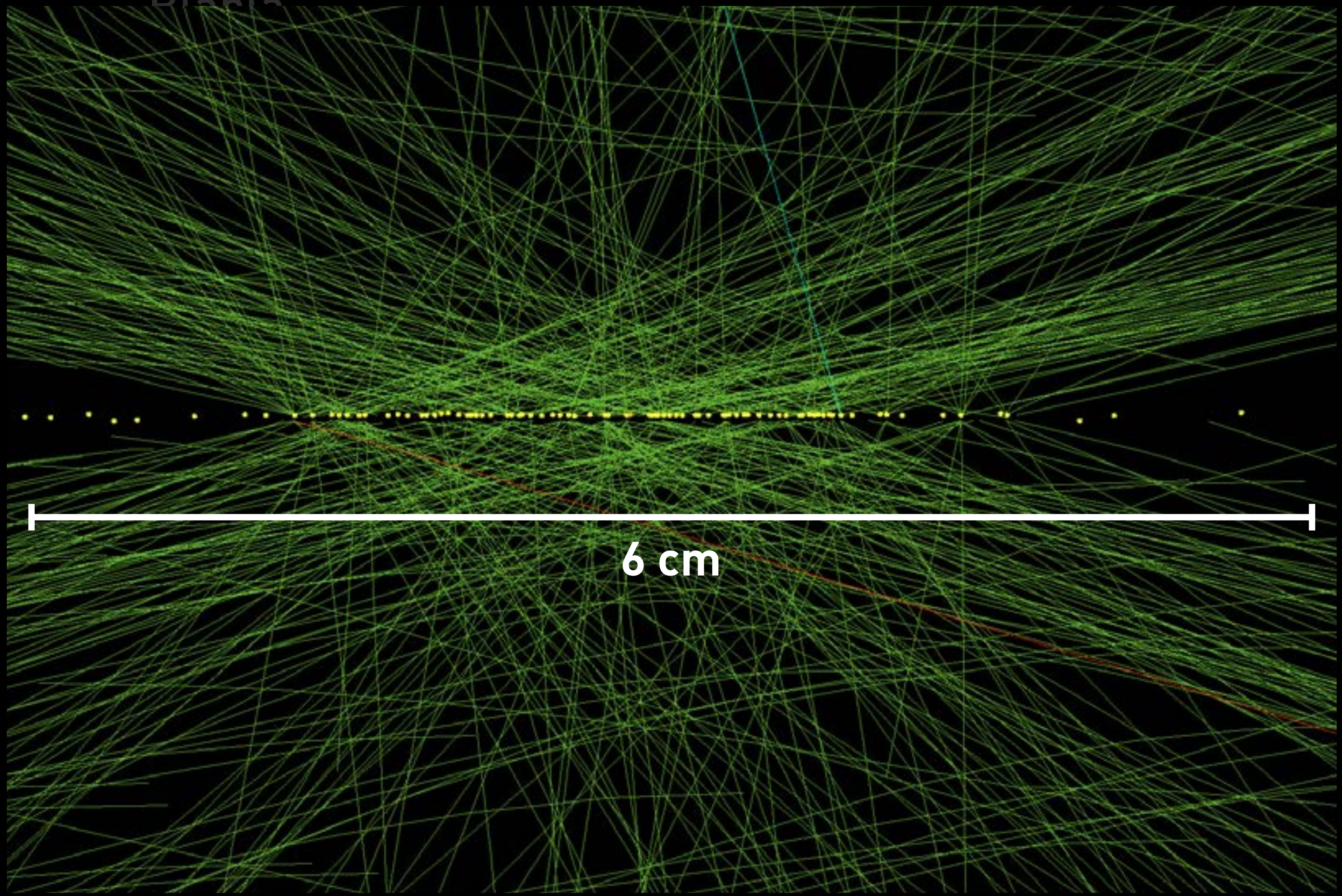
- $\times$10 data size
- $\times$3 collisions/s



2022 - 2025	2026 - 2028	2029 - 2038
LHC (TODAY!)	MAJOR UPGRADE	HL-LHC
Run 3		Run 4+5

LHC

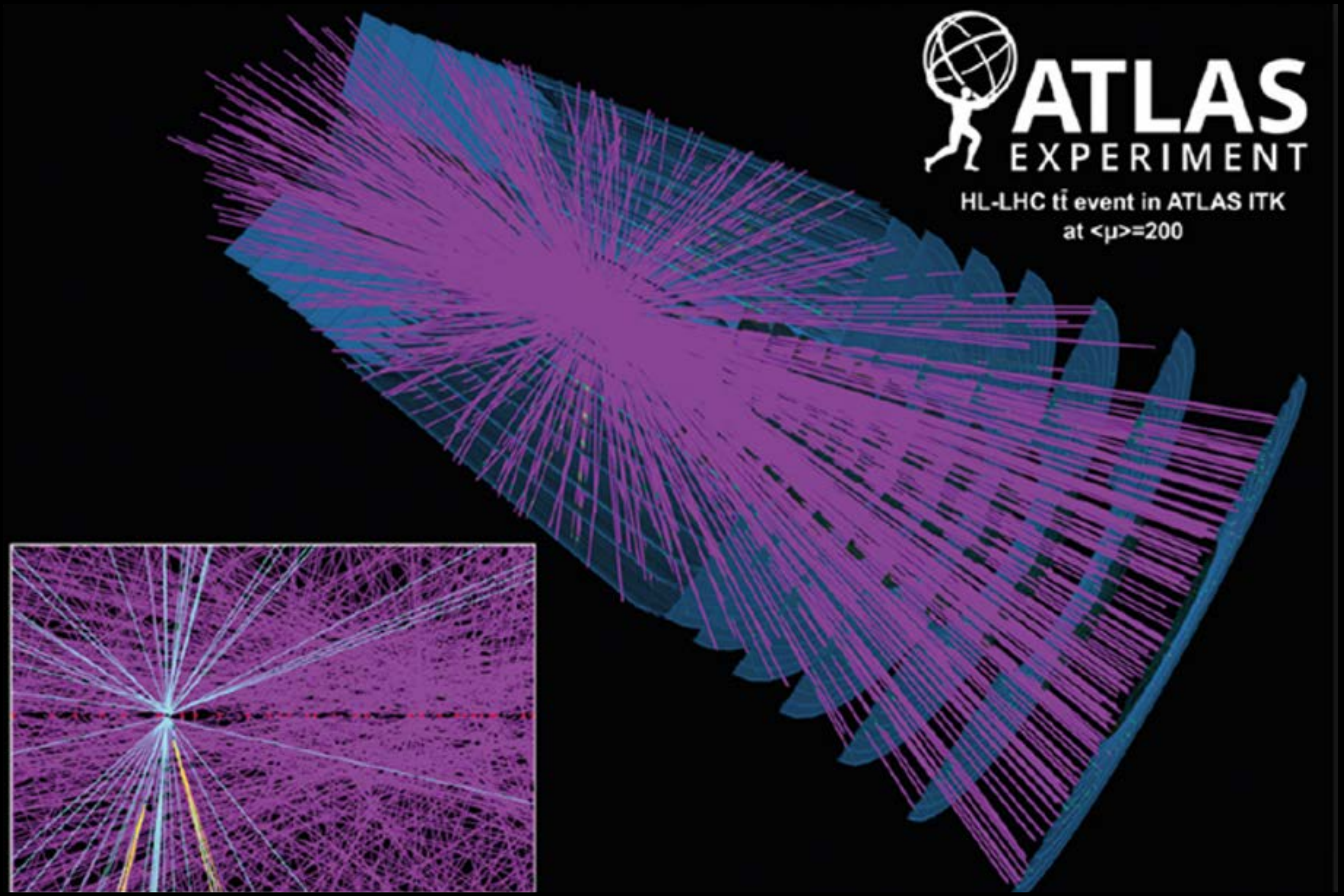
78 vertices
(average 60)



Run 3

High Luminosity LHC

200 vertices
(average 140)



Run 4+5

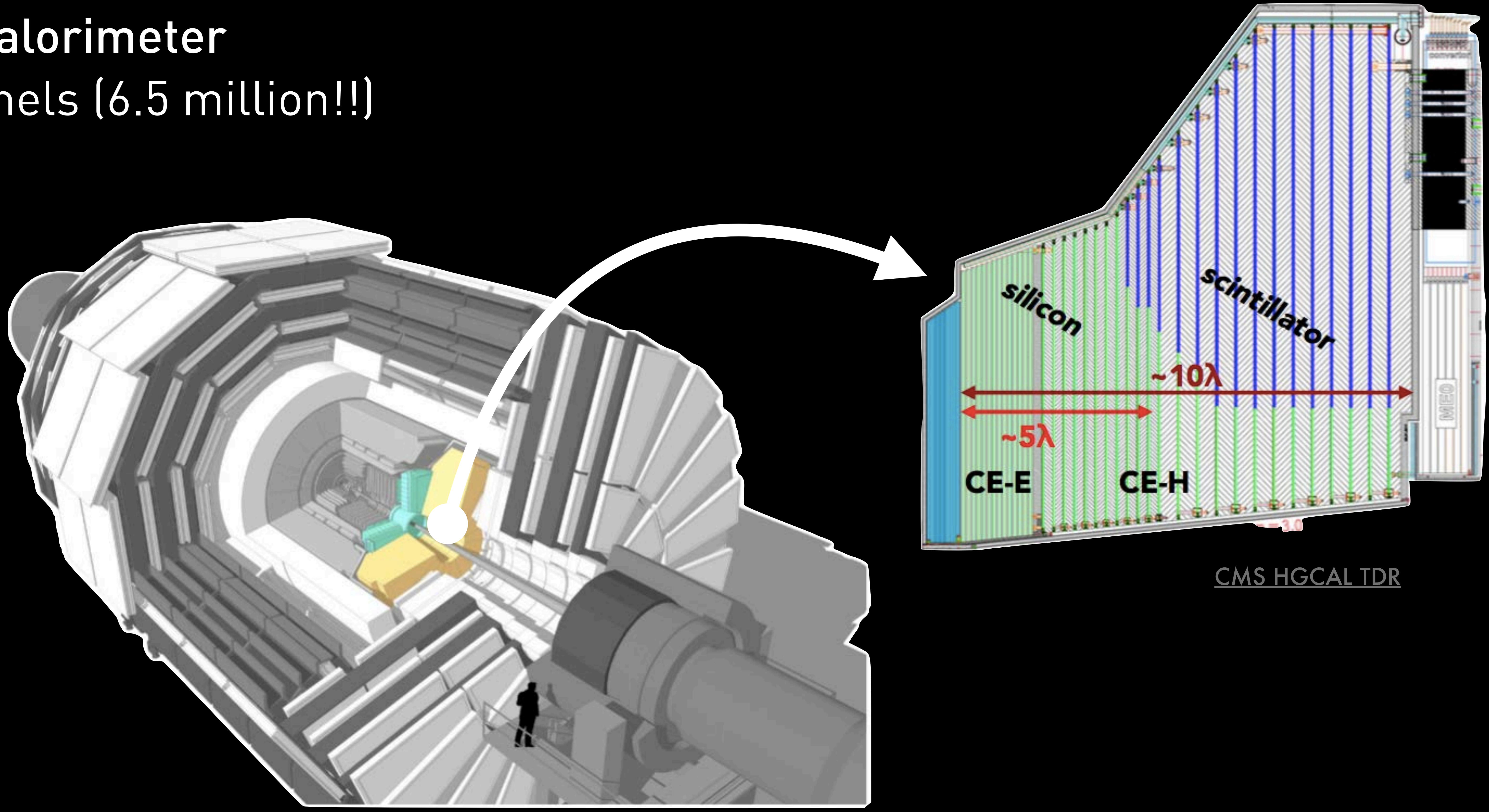
Maintain physics acceptance → better detectors

CMS High Granularity (endcap) calorimeter

- X20 times more readout channels (6.5 million!!)

More collisions

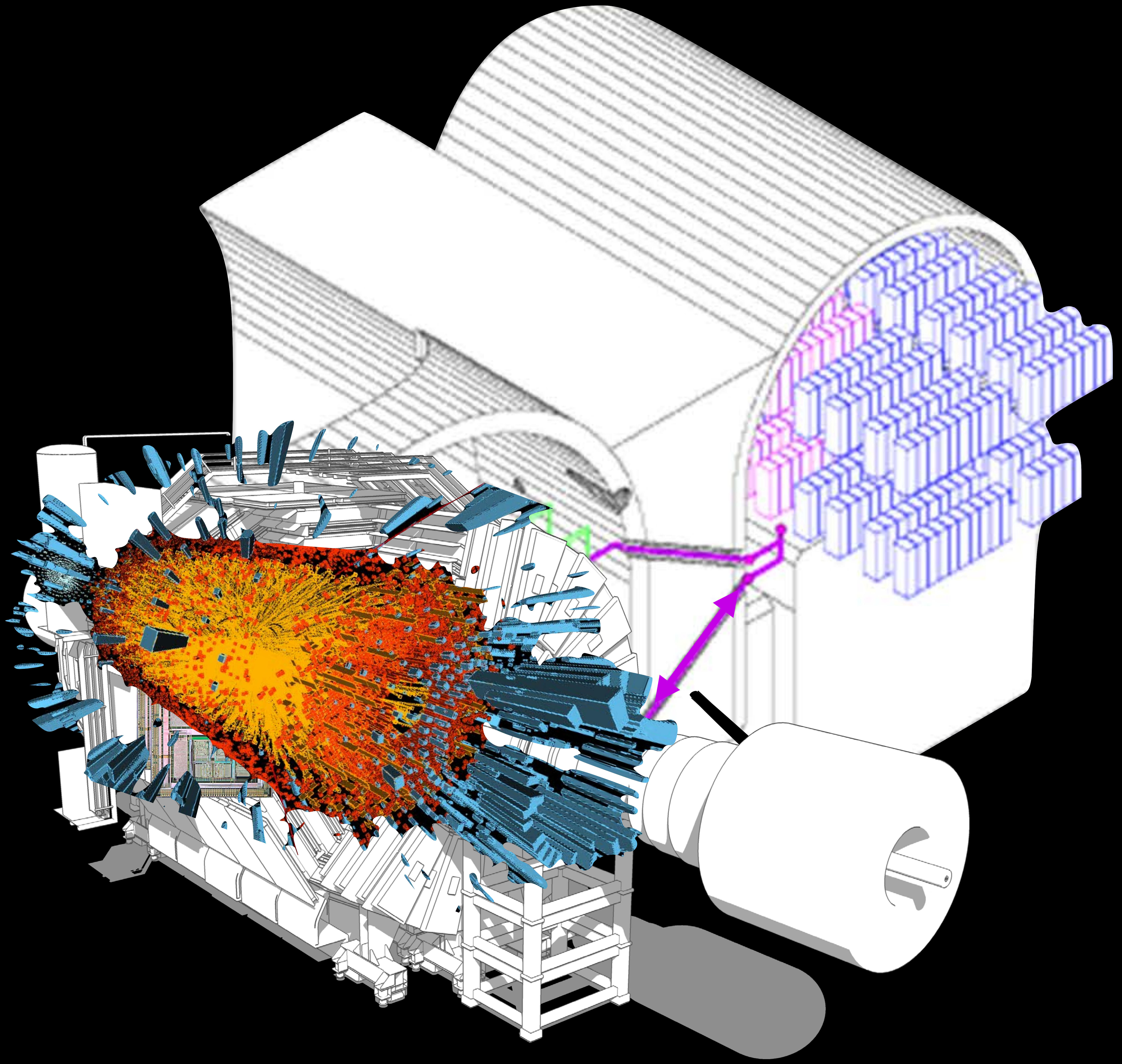
More readout channels



CMS HGCAL TDR

HL-LHC Level-1:

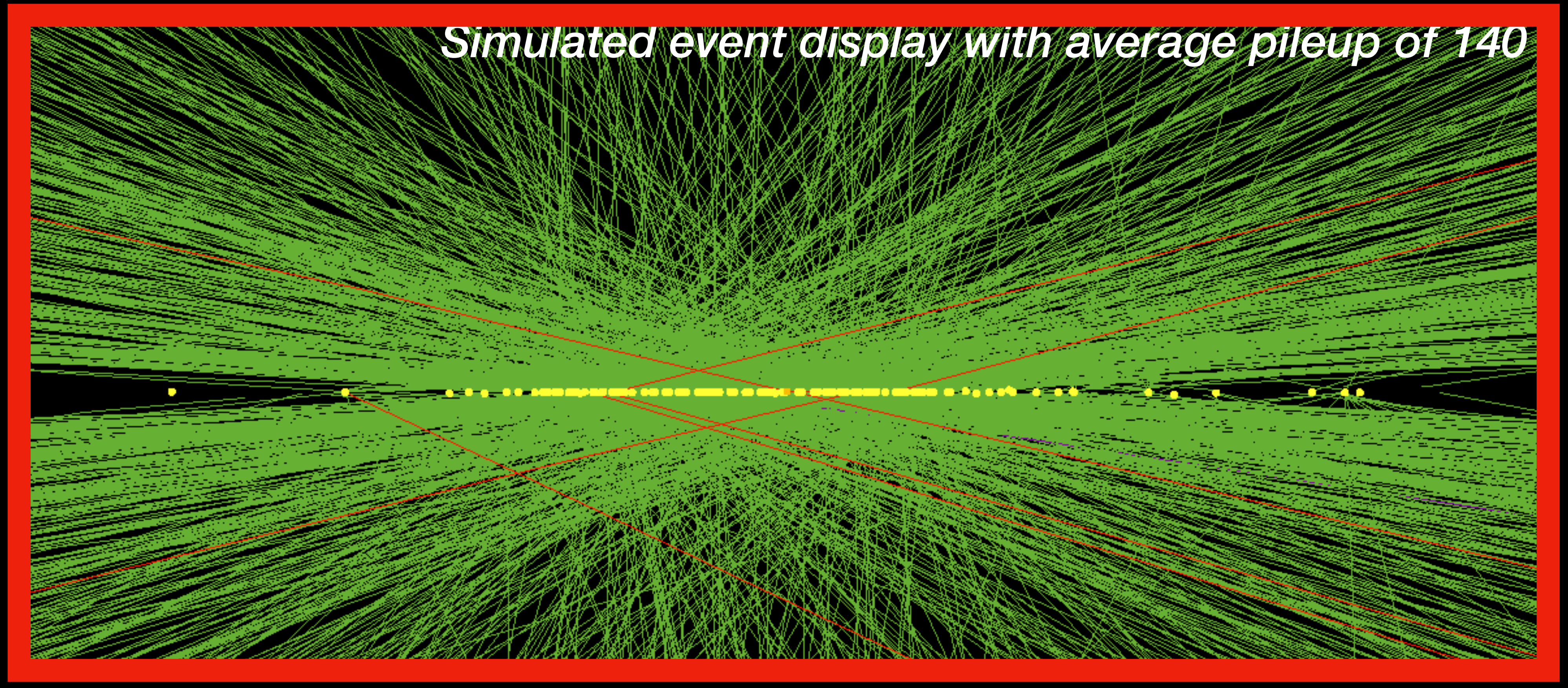
Complete re-design of Level-1

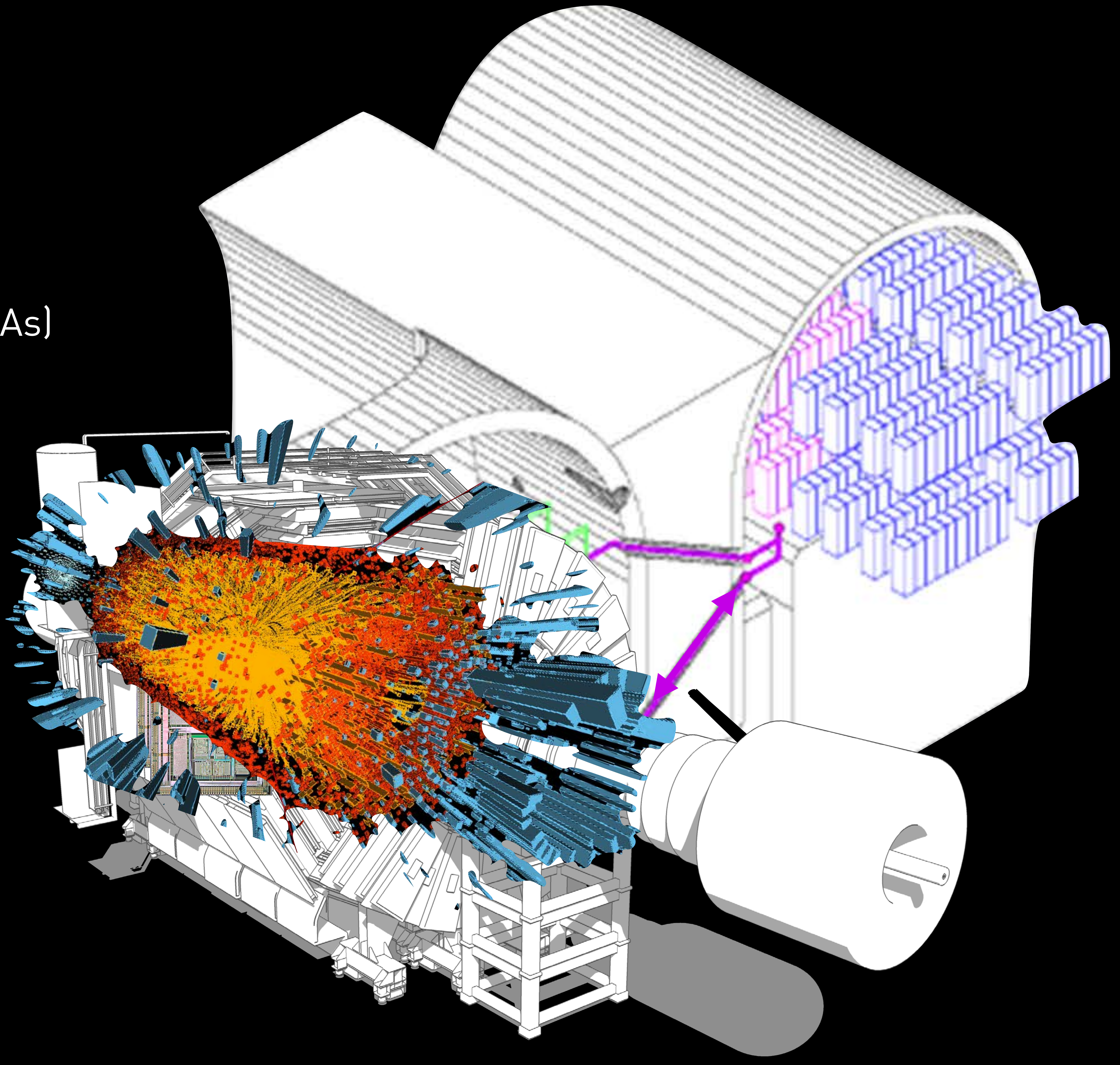


HL-LHC Level-1:

Complete re-design of Level-1

- Charged particle tracks (6.4 Tb/s, 200 FPGAs)

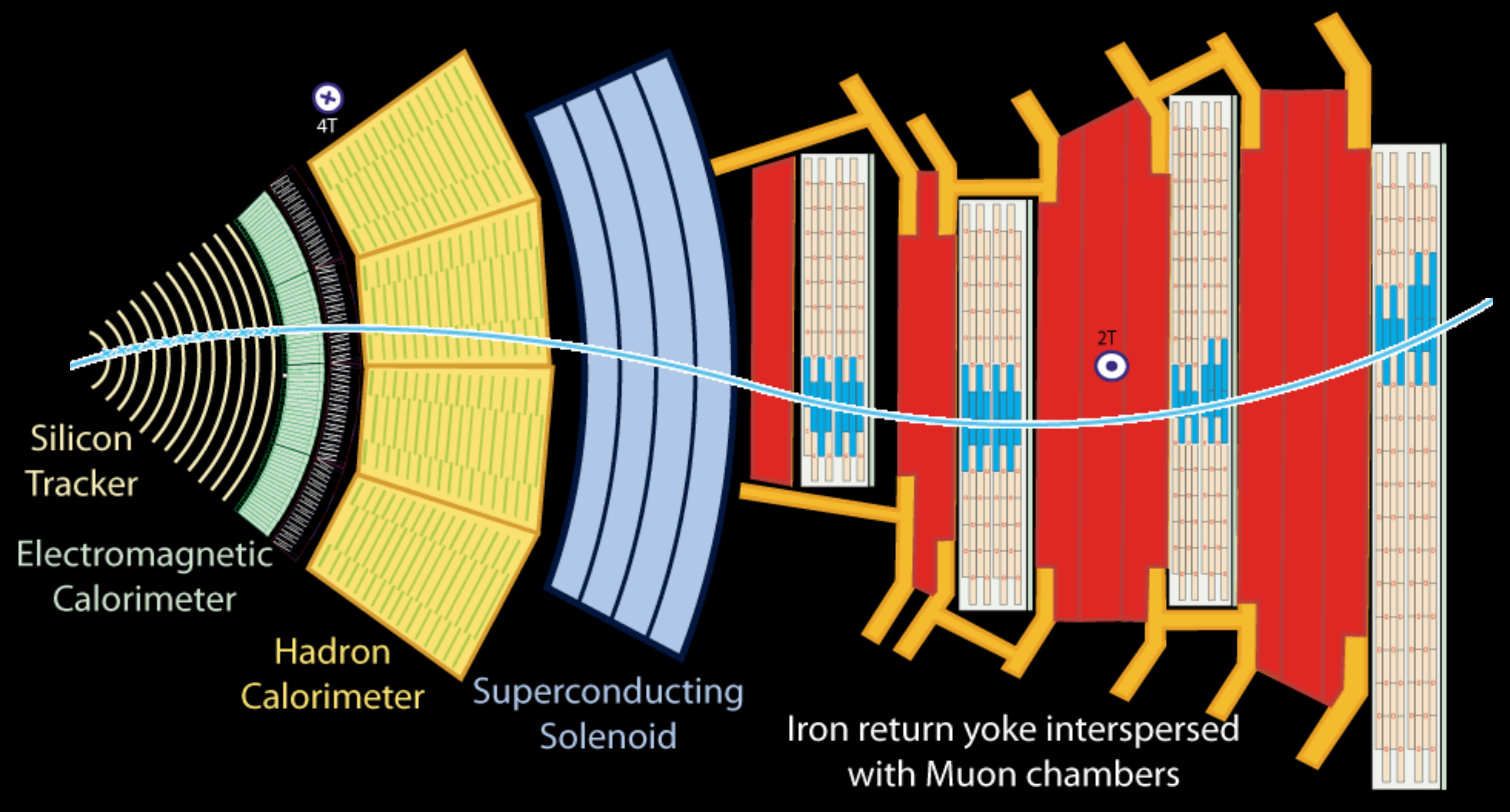


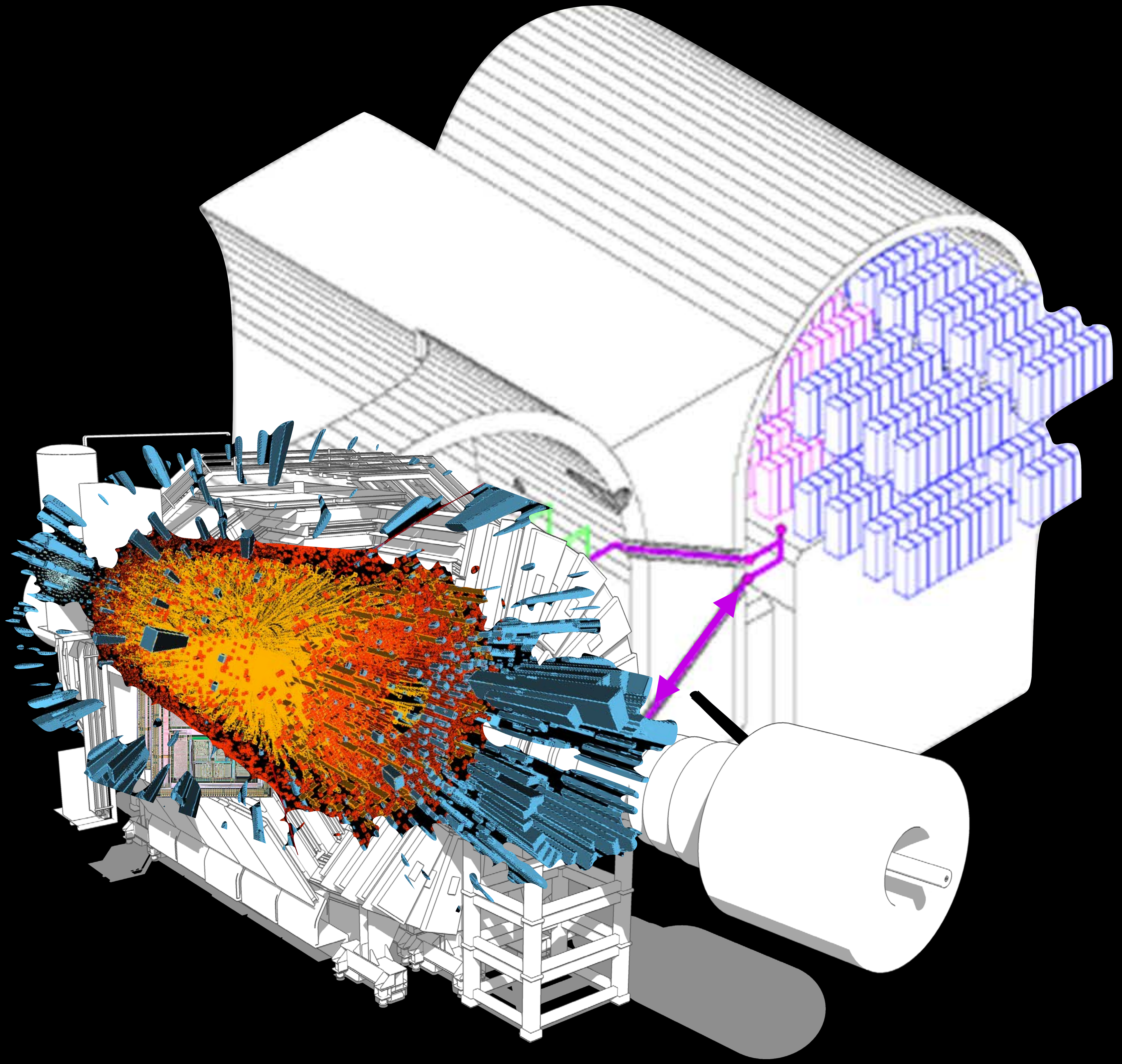


HL-LHC Level-1:

Complete re-design of Level-1

- Charged particle tracks
- Particle Flow (40 FPGAs)

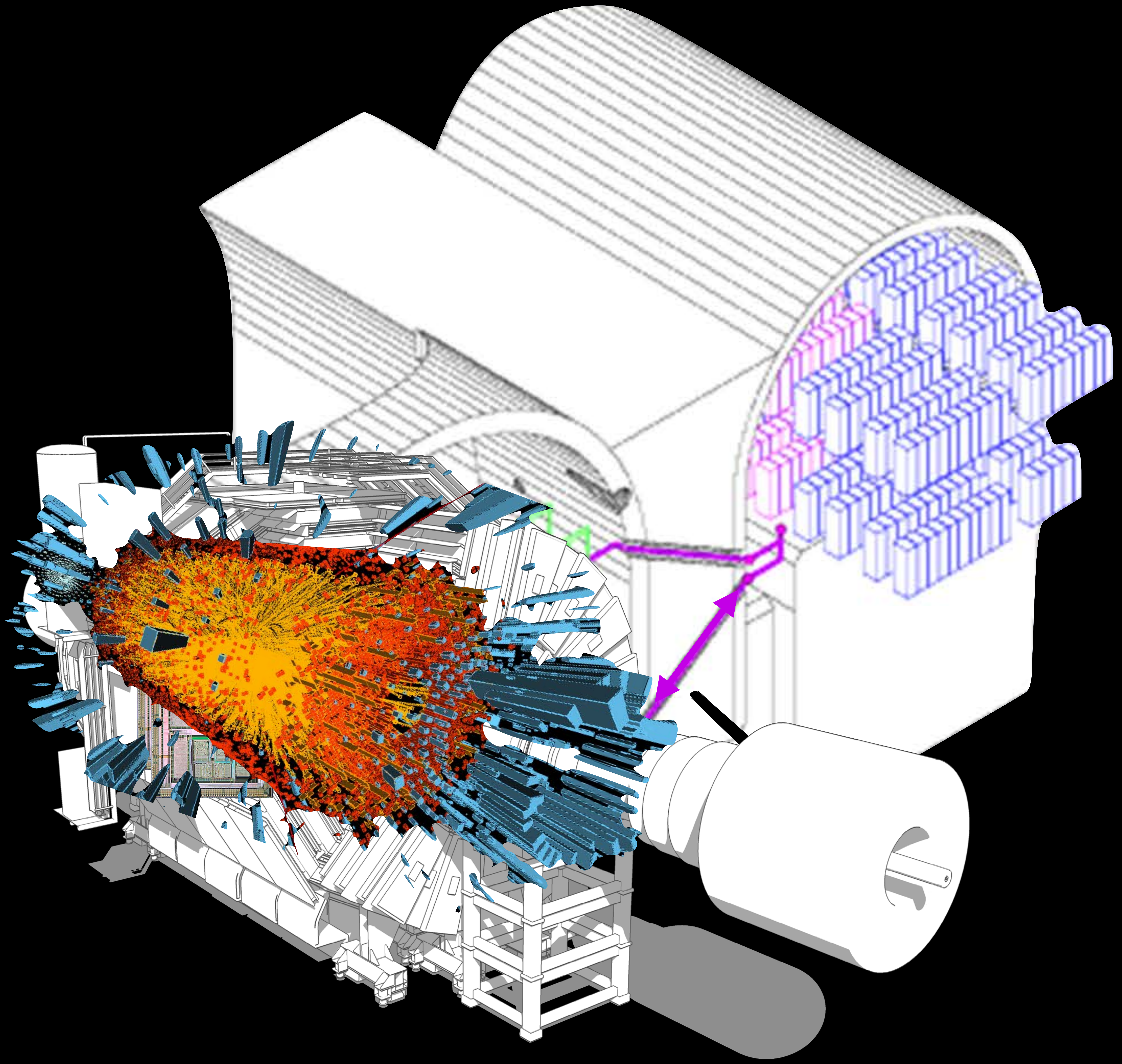




HL-LHC Level-1:

Complete re-design of Level-1

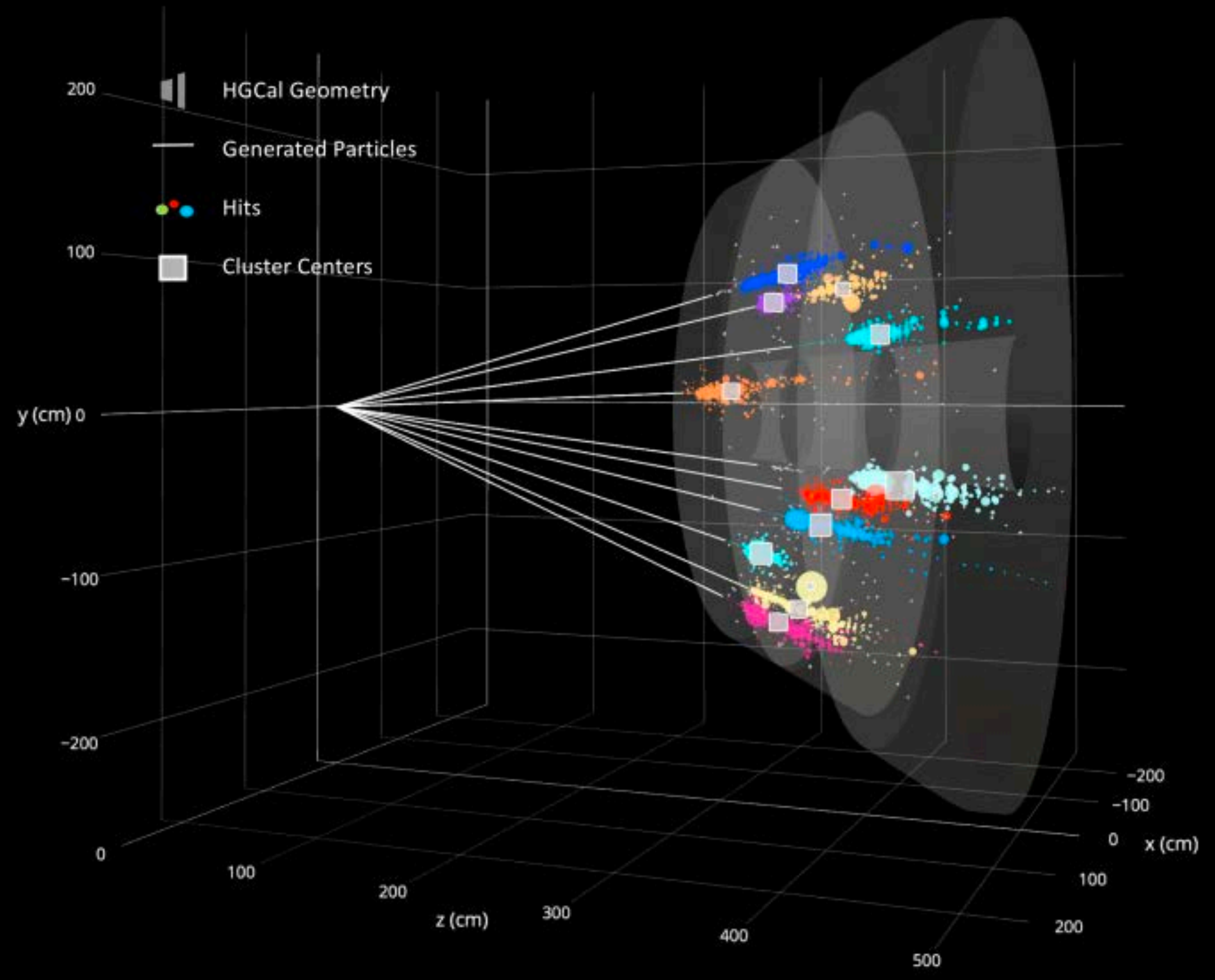
- Charged particle tracks
- Particle Flow (40 FPGAs)

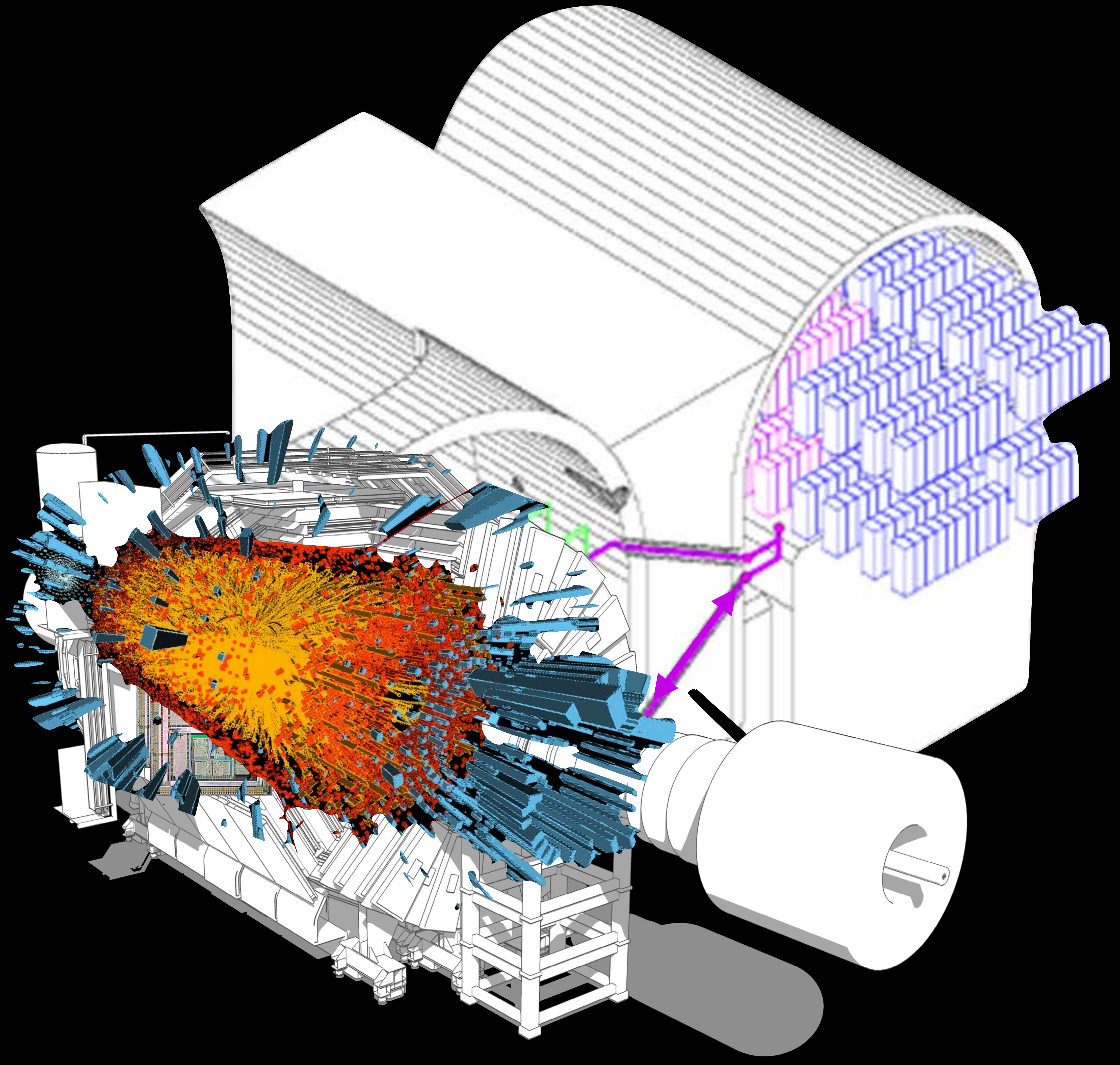


HL-LHC Level-1:

Complete re-design of Level-1

- Charged particle tracks
- Particle Flow
- HGCal (4 Tb/s, 200 FPGAs)



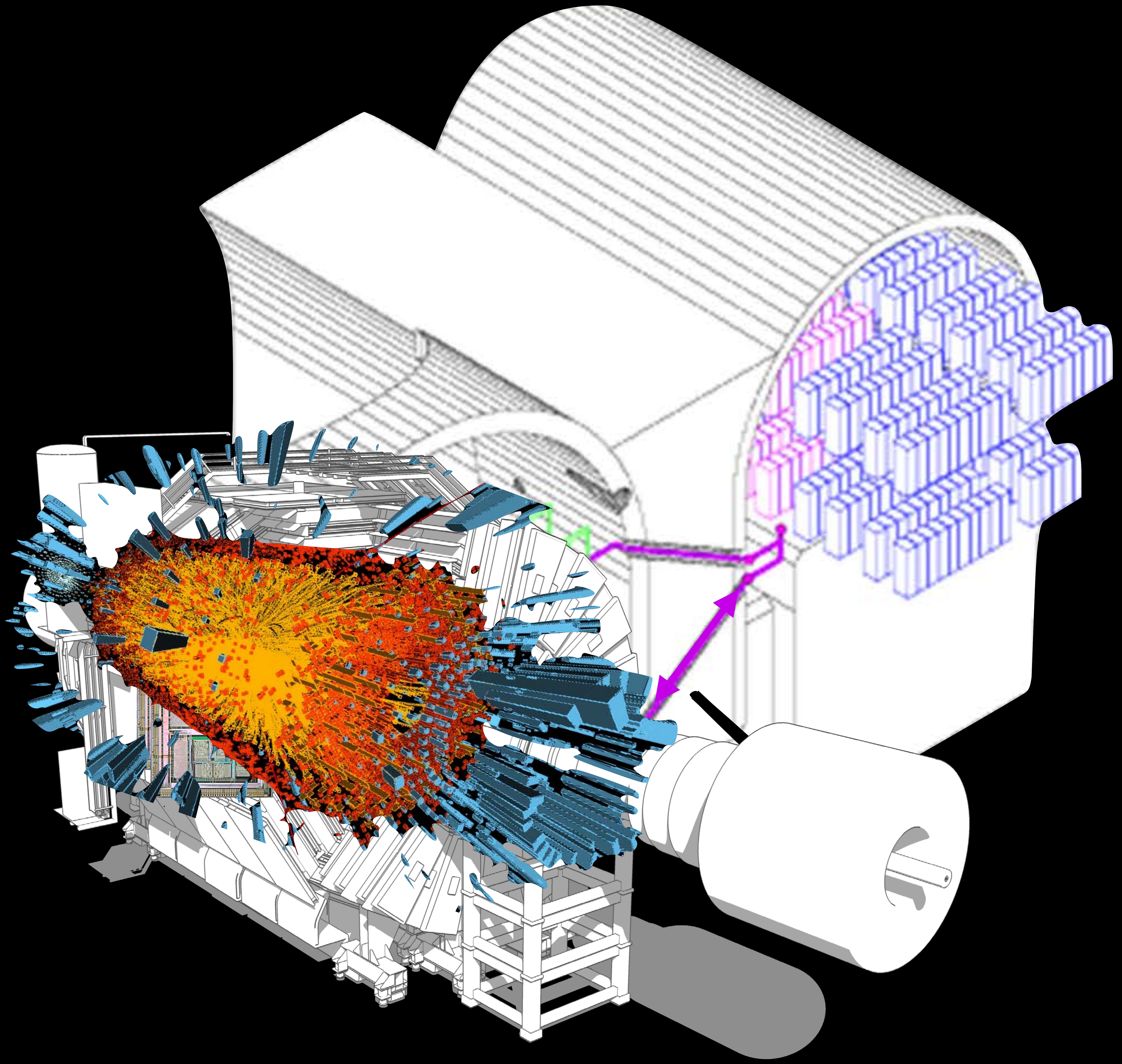


HL-LHC Level-1:

Complete re-design of Level-1
- Charged particle tracks
- Particle Flow
- HGCal

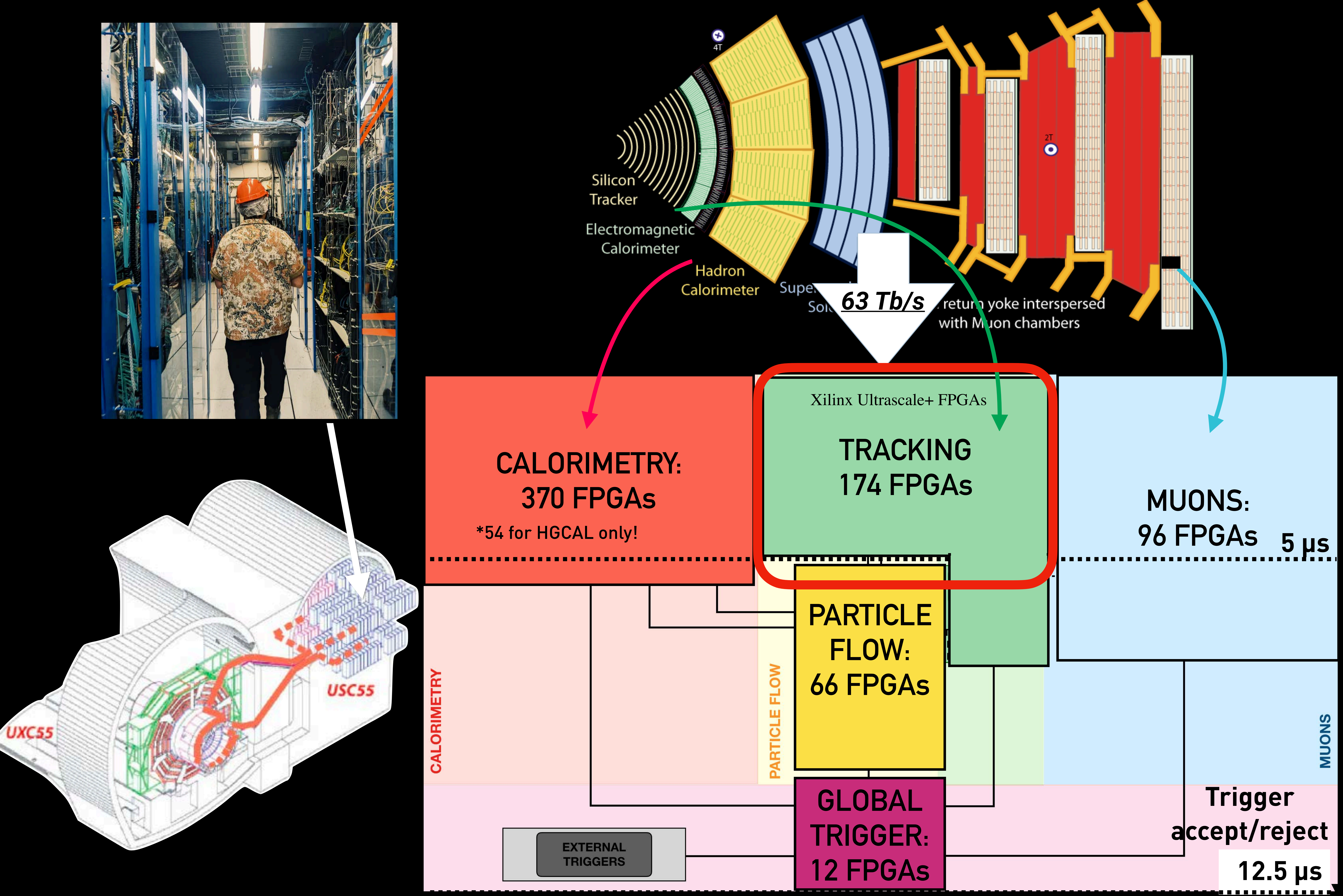
Input data
- 2 Tb/s → **63 Tb/s**

Latency
- 4 μs → **12 μs**

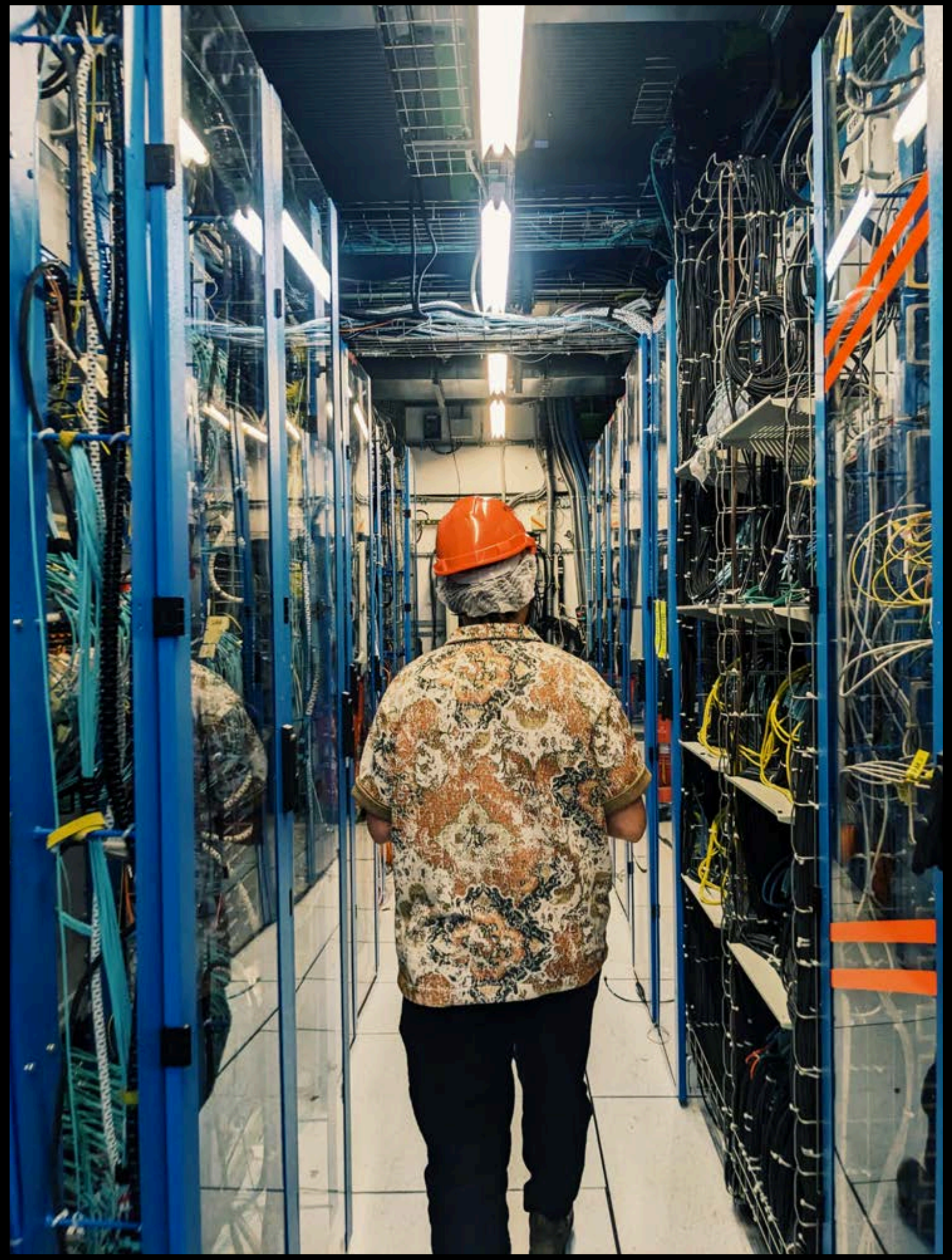


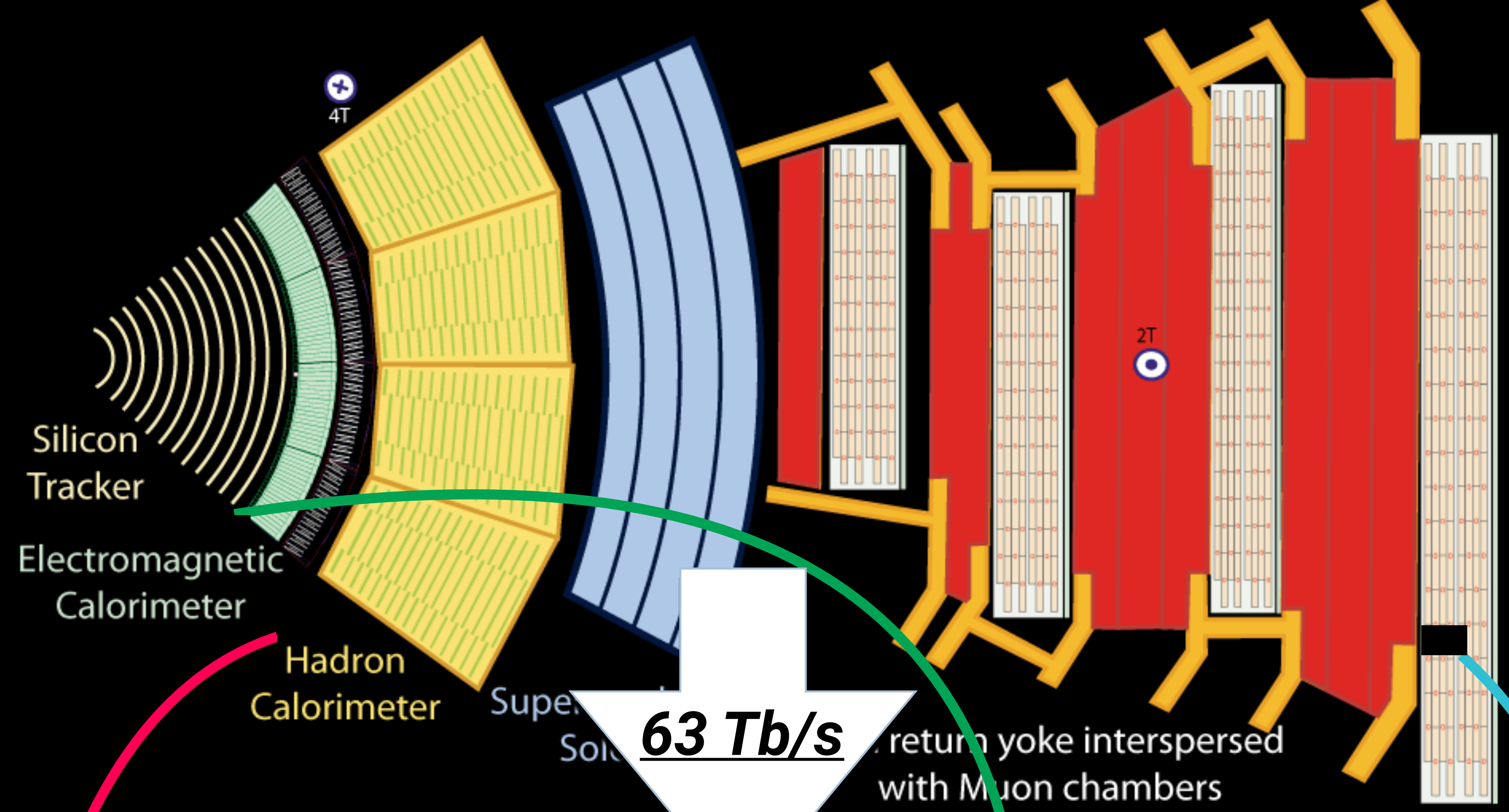
Extremely high data complexity,

Extremely little time

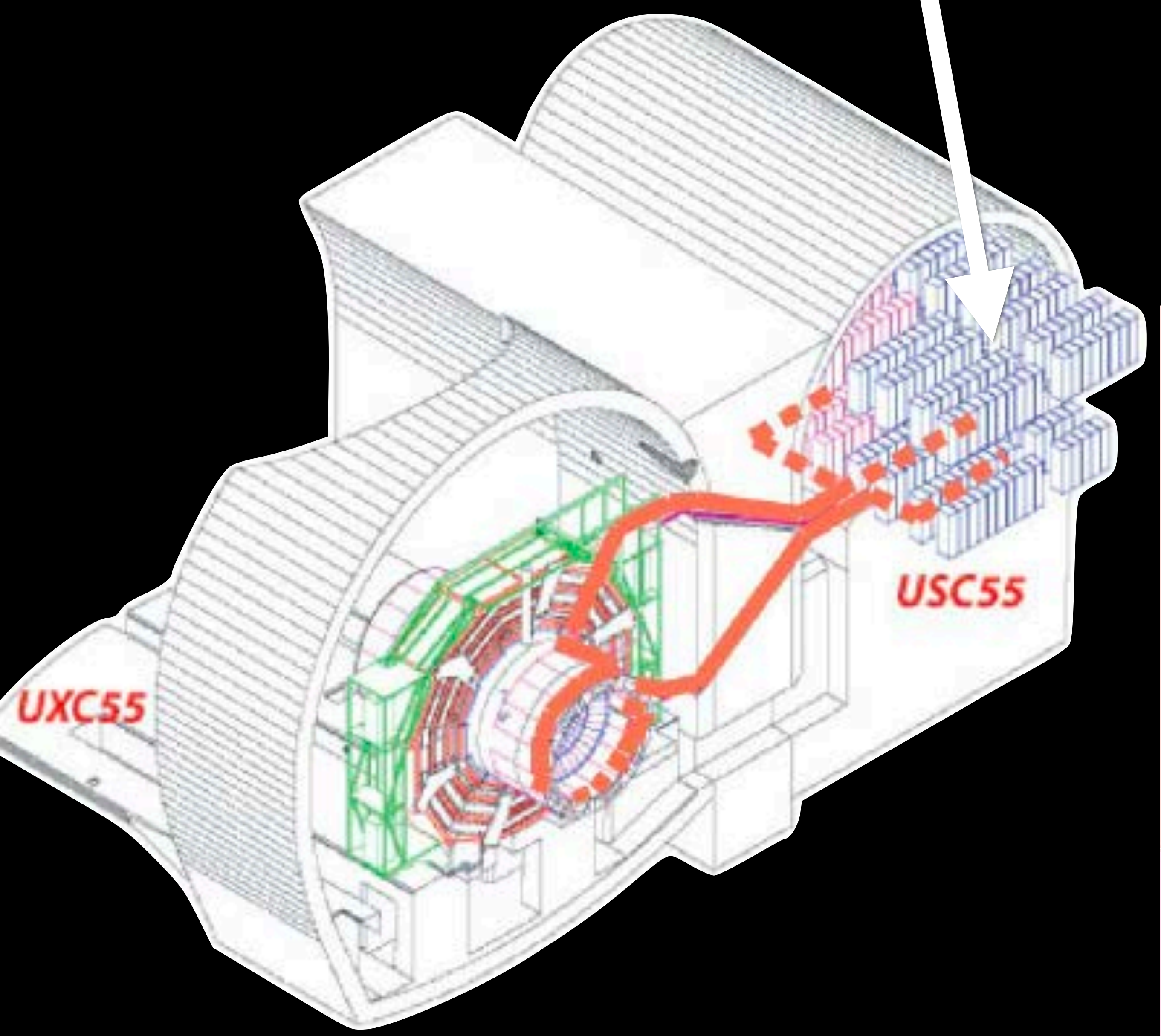


Silicon Tracker
Electromagnetic Calorimeter
Hadron Calorimeter
4T
2T
63 Tb/s
with Muon chambers
Xilinx Ultrascale+ FPGAs
CALORIMETRY: 370 FPGAs
*54 for HGCAL only!
TRACKING 174 FPGAs
MUONS: 96 FPGAs
5 μs
PARTICLE FLOW: 66 FPGAs
GLOBAL TRIGGER: 12 FPGAs
EXTERNAL TRIGGERS
Trigger accept/reject
12.5 μs
CALORIMETRY
PARTICLE FLOW
MUONS
UXC55
USC55

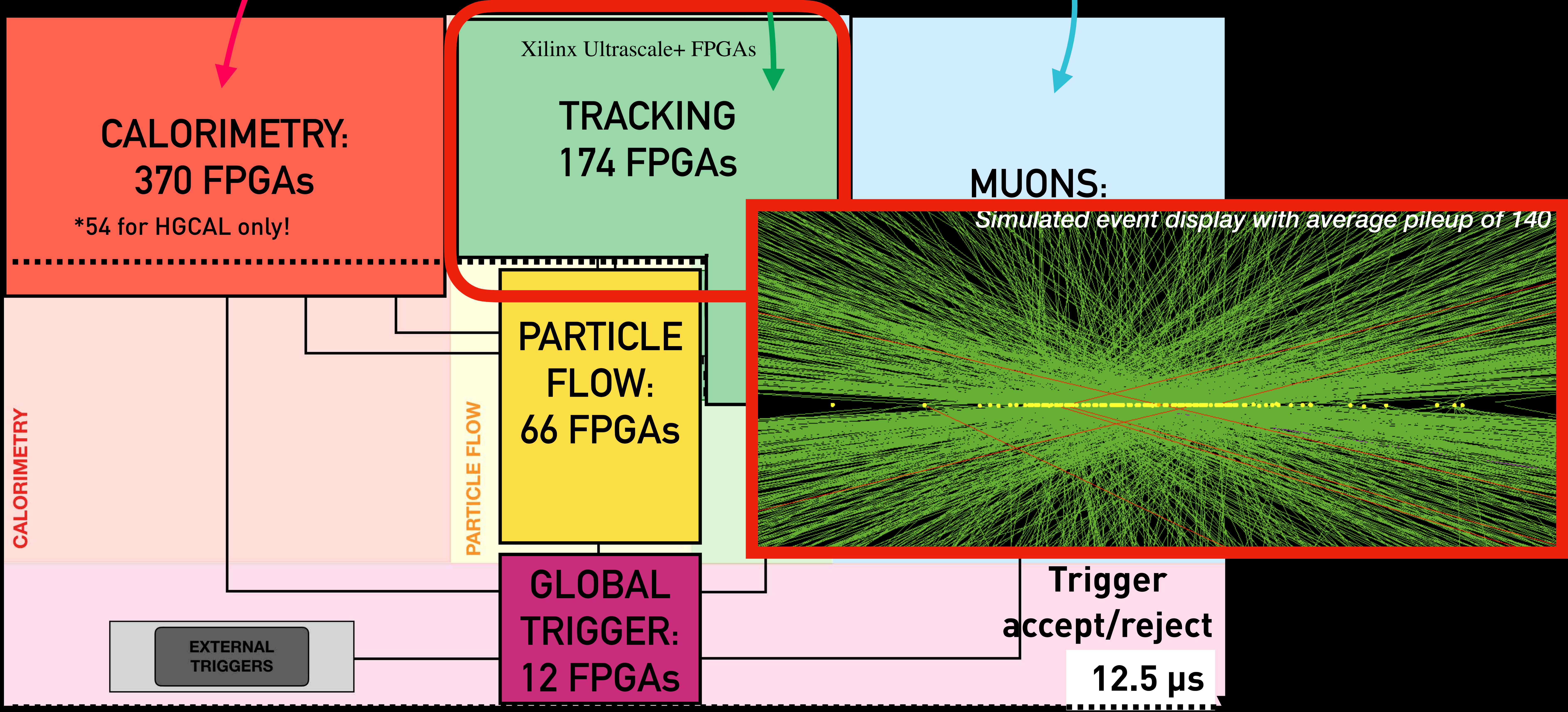




4T
Silicon
Tracker
Electromagnetic
Calorimeter
Hadron
Calorimeter
63 Tb/s
2T

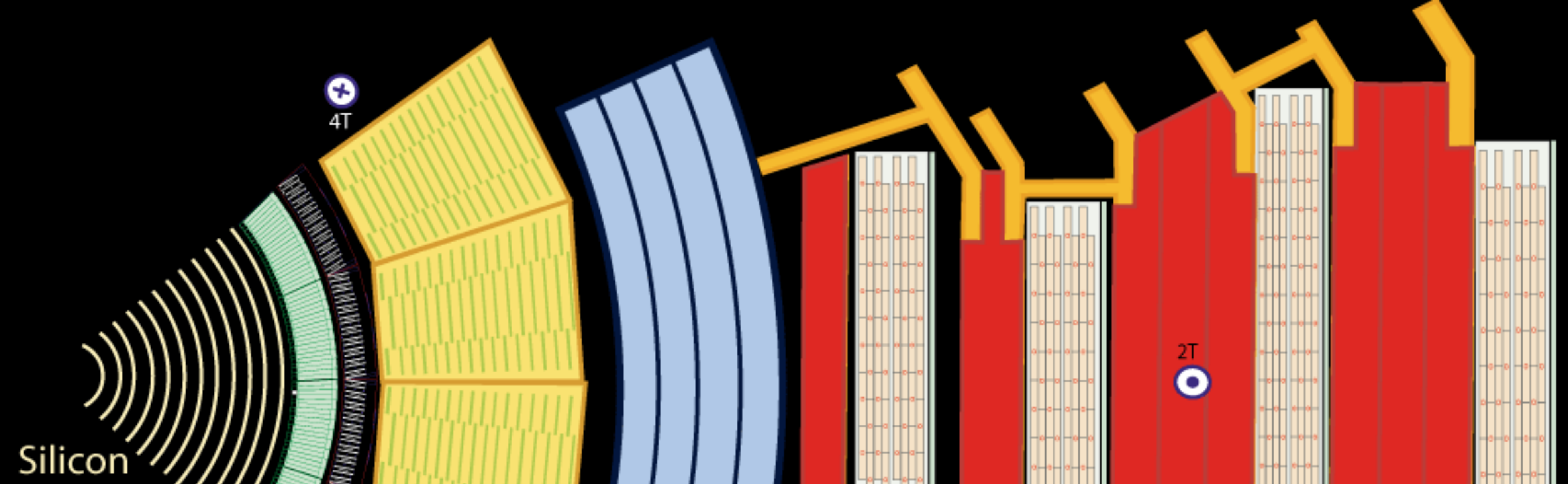


USC55
UXC55



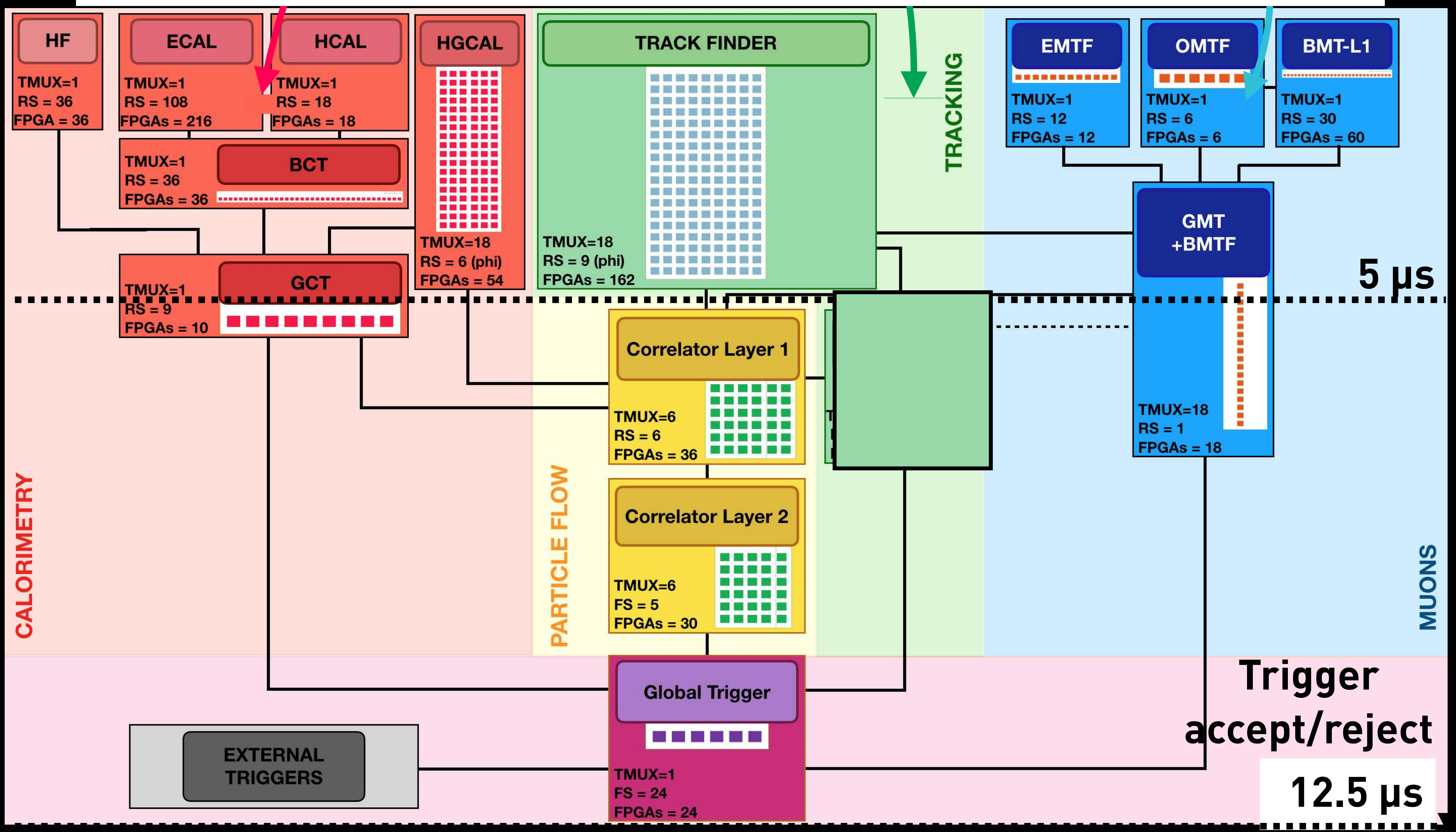
CALORIMETRY:
370 FPGAs
*54 for HGCAL only!
Xilinx Ultrascale+ FPGAs
TRACKING
174 FPGAs
MUONS:
PARTICLE
FLOW:
66 FPGAs
CALORIMETRY
PARTICLE FLOW
GLOBAL
TRIGGER:
12 FPGAs
EXTERNAL
TRIGGERS
Trigger
accept/reject
12.5 μs
Simulated event display with average pileup of 140

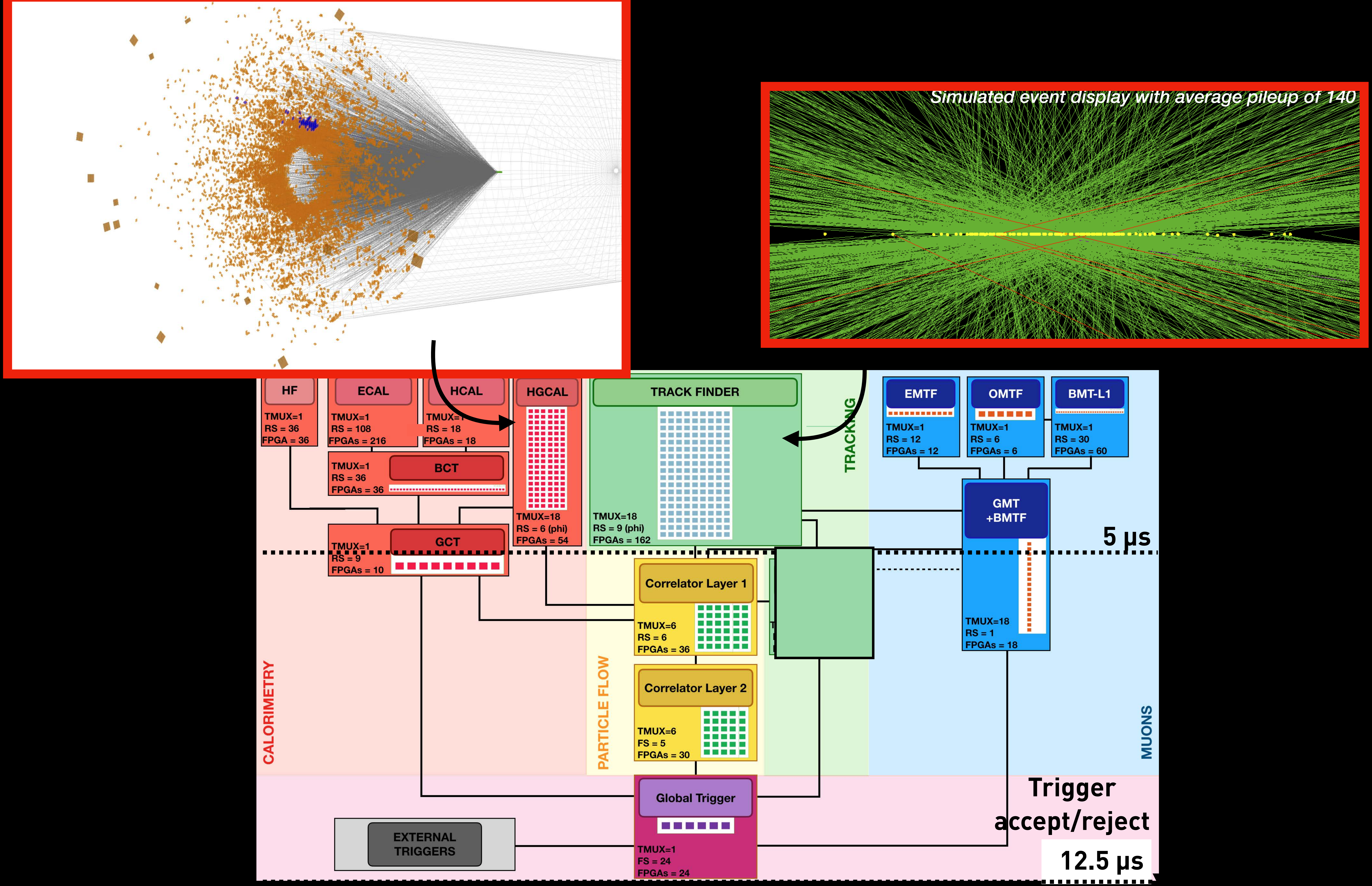
BRAND NEW SYSTEM, NOT YET BUILT!
The time to design algorithms is now
Silicon Tracker
Electromagnetic Calorimeter
Hadron Calorimeter
63 Tb/s
with Muon chambers
Xilinx Ultrascale+ FPGAs
event display with average pileup of 140
UXC55
USC55
CALORIMETRY
PARTICLE FLOW
PARTICLE FLOW: 66 FPGAs
GLOBAL TRIGGER: 12 FPGAs
EXTERNAL TRIGGERS
Trigger accept/reject
12.5 μs



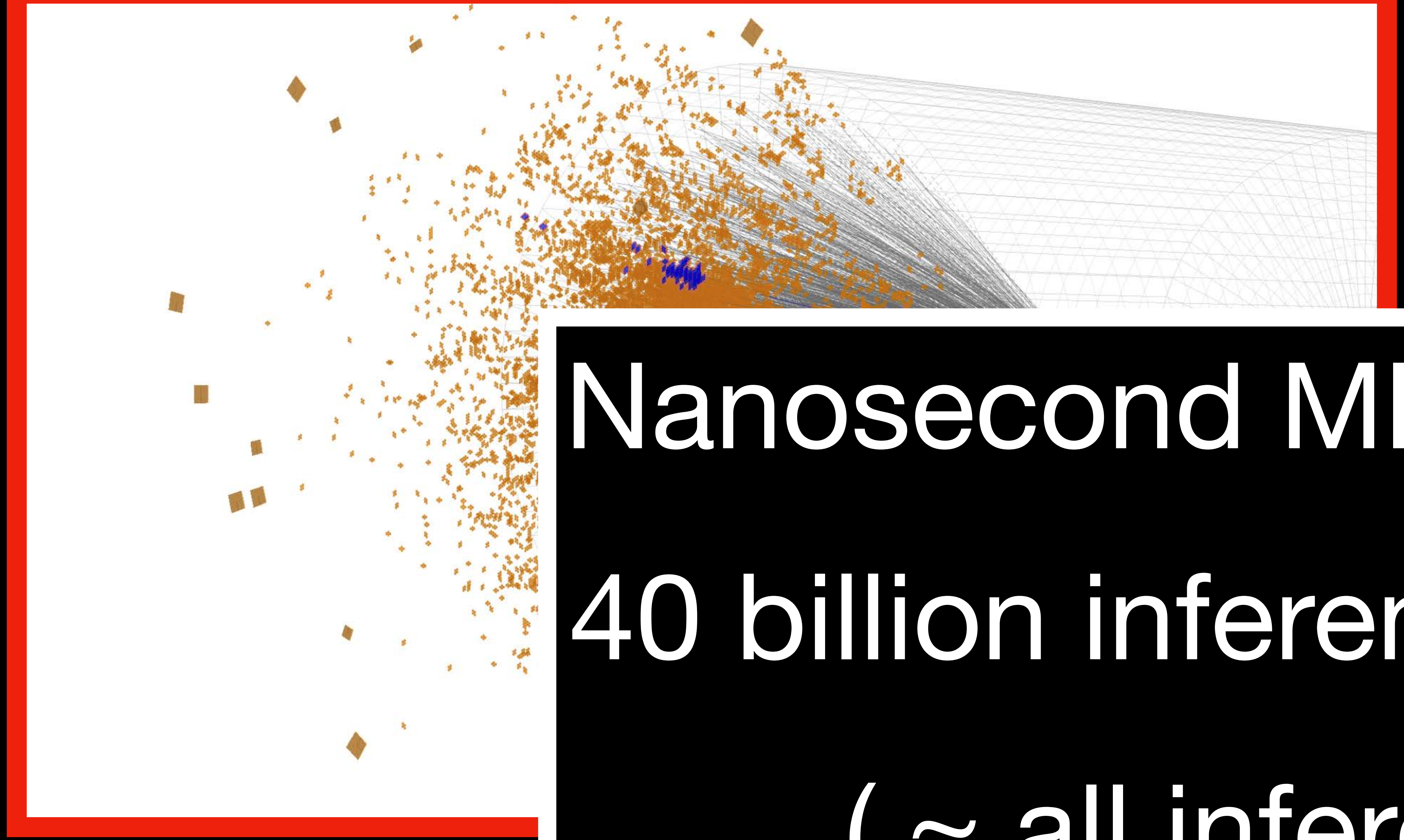
12 microseconds latency

Processing 5% of internet traffic





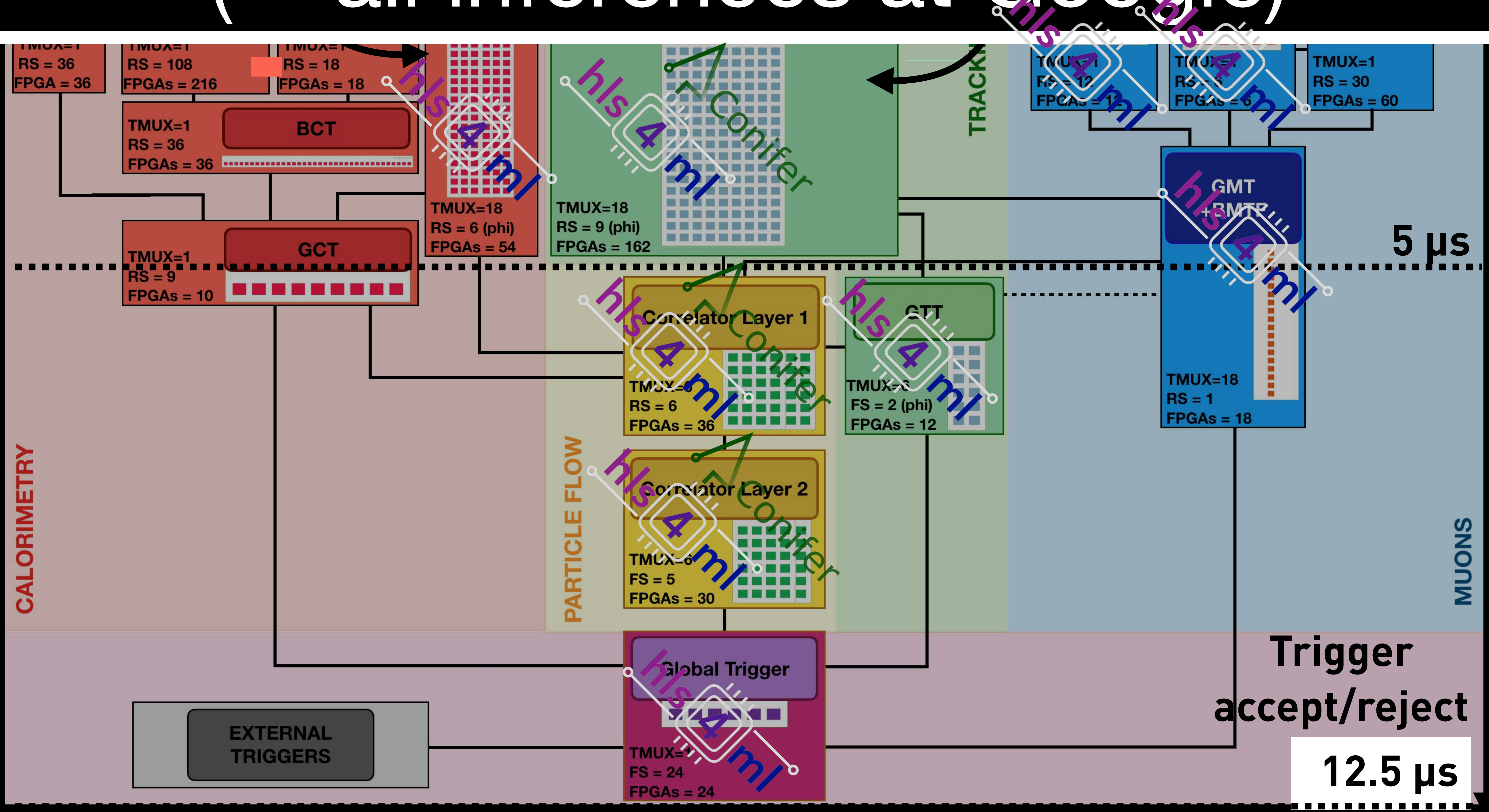
Simulated event display with average pileup of 140
HF
TMUX=1
RS = 36
FPGA = 36
ECAL
TMUX=1
RS = 108
FPGAs = 216
HCAL
TMUX=1
RS = 18
FPGAs = 18
HGCAL
TMUX=18
RS = 6 (phi)
FPGAs = 54
BCT
TMUX=1
RS = 36
FPGAs = 36
GCT
TMUX=1
RS = 9
FPGAs = 10
TRACK FINDER
TMUX=18
RS = 9 (phi)
FPGAs = 162
TRACKING
EMTF
TMUX=1
RS = 12
FPGAs = 12
OMTF
TMUX=1
RS = 6
FPGAs = 6
BMT-L1
TMUX=1
RS = 30
FPGAs = 60
GMT
+BMTF
TMUX=18
RS = 1
FPGAs = 18
5 μs
Correlator Layer 1
TMUX=6
RS = 6
FPGAs = 36
Correlator Layer 2
TMUX=6
FS = 5
FPGAs = 30
CALORIMETRY
PARTICLE FLOW
MUONS
Global Trigger
TMUX=1
FS = 24
FPGAs = 24
EXTERNAL
TRIGGERS
Trigger
accept/reject
12.5 μs

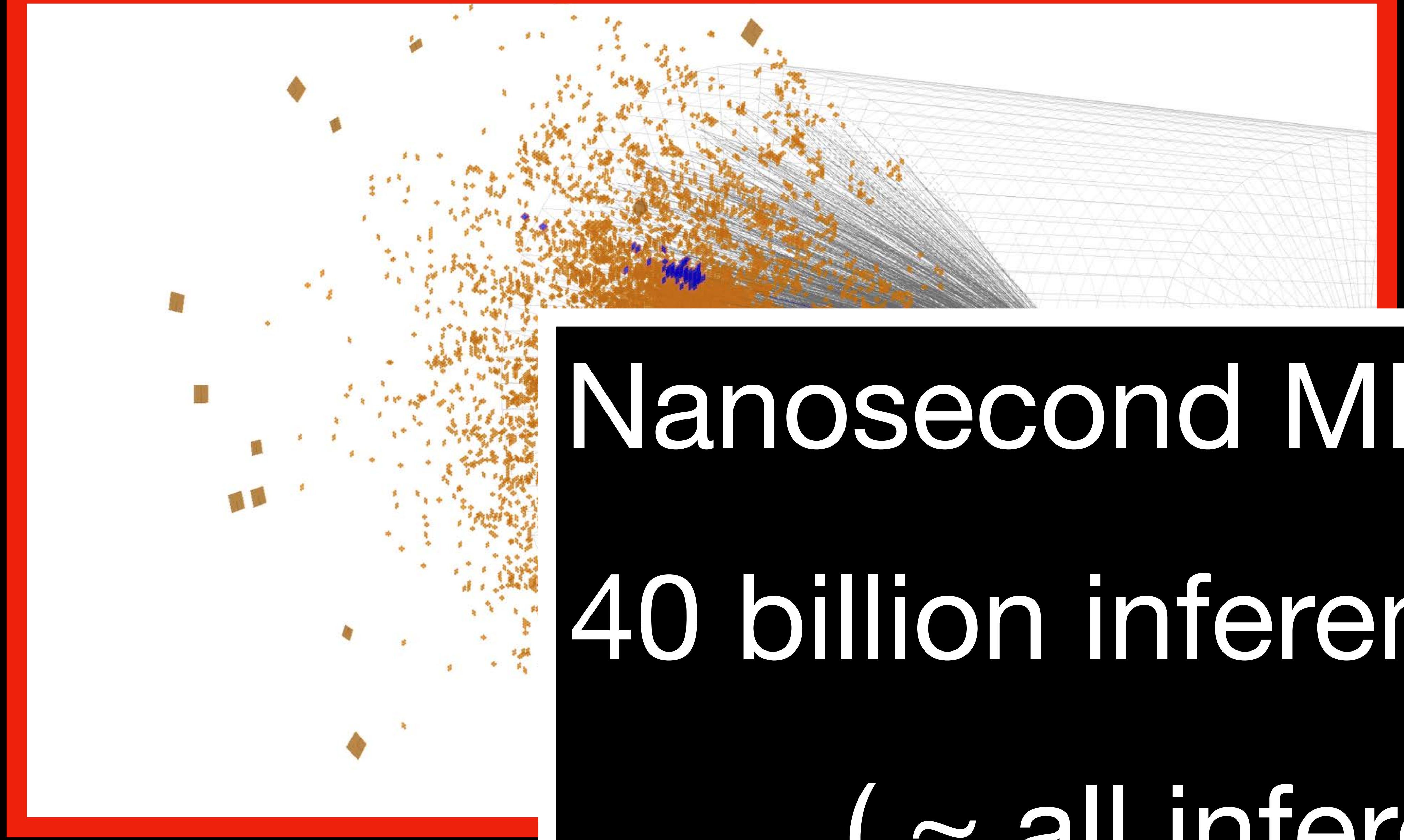


Nanosecond ML inference on FPGAs!

40 billion inferences/s during HL-LHC

(≈ all inferences at Google)

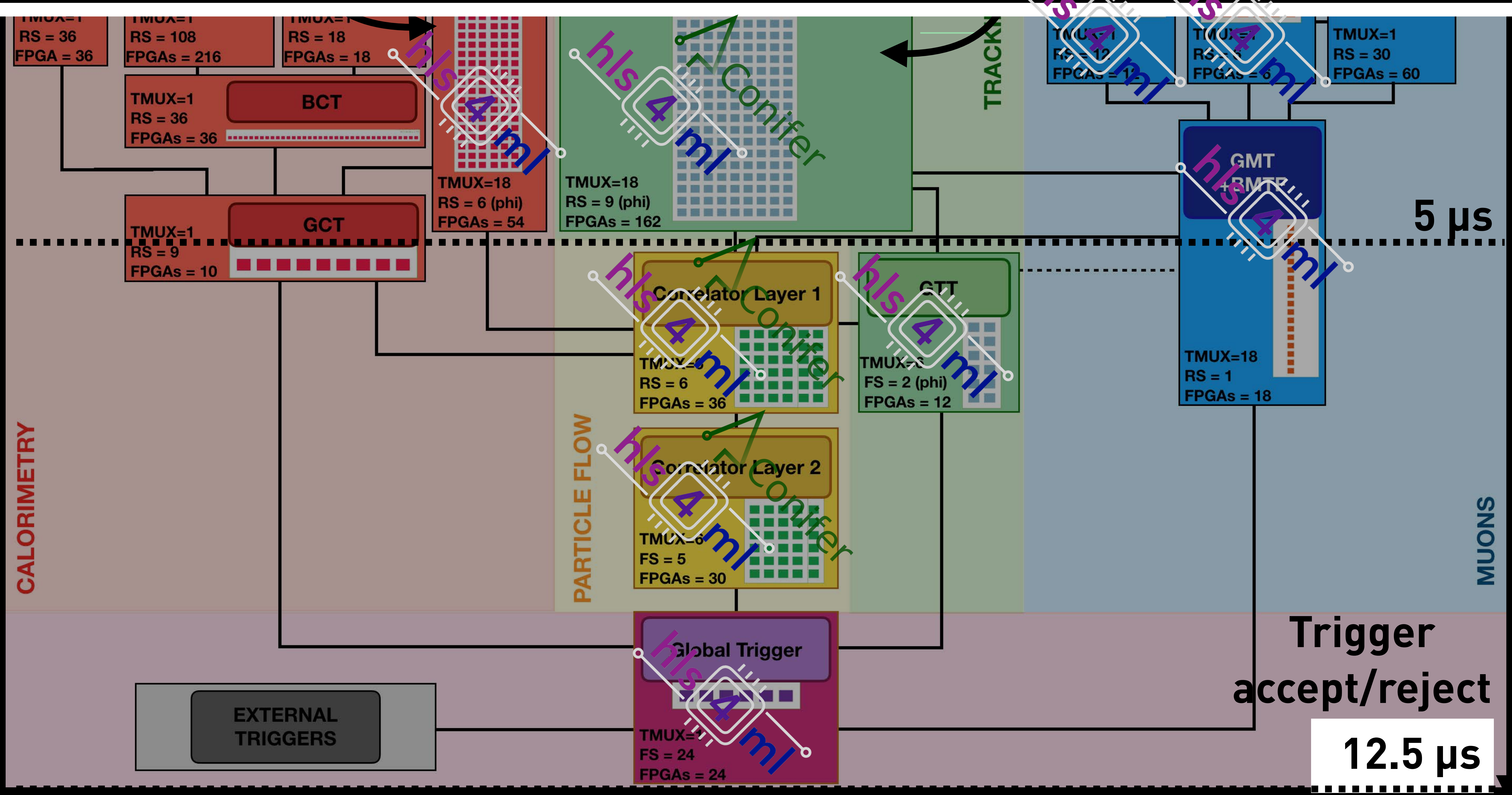




Nanosecond ML inference on FPGAs!

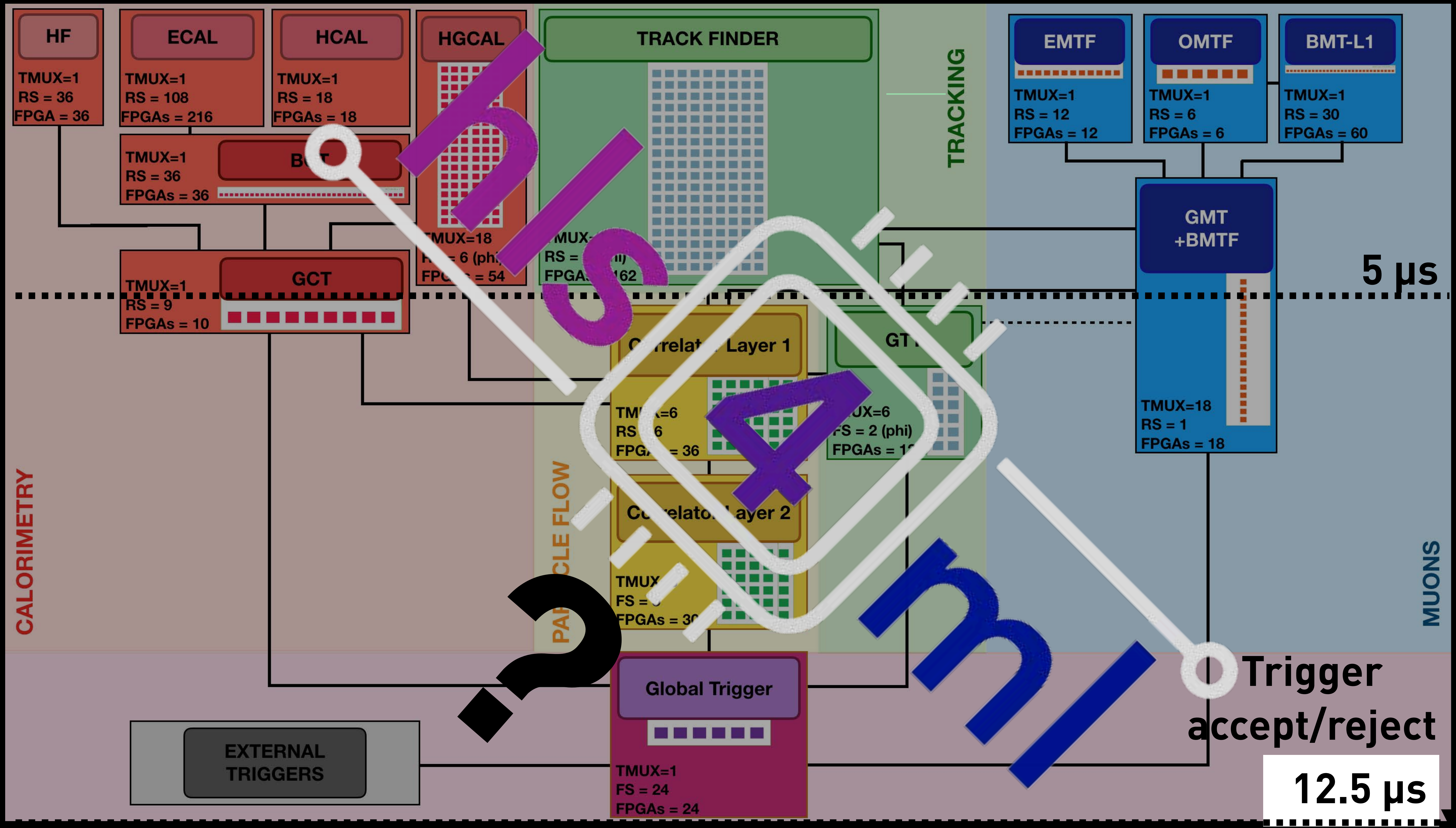
40 billion inferences/s during HL-LHC

(≈ all inferences at Google)



Conifer hls4ml

HEP developed libraries for fast ML on FPGAs


HF
TMUX=1
RS = 36
FPGA = 36
ECAL
TMUX=1
RS = 108
FPGAs = 216
HCAL
TMUX=1
RS = 18
FPGAs = 18
HGCAL
TMUX=18
TMUX=1
RS = 36
FPGAs = 36
GCT
TMUX=1
RS = 9
FPGAs = 10
TRACK FINDER
TRACKING
EMTF
TMUX=1
RS = 12
FPGAs = 12
OMTF
TMUX=1
RS = 6
FPGAs = 6
BMT-L1
TMUX=1
RS = 30
FPGAs = 60
GMT
+BMTF
5 μs
TMUX=18
RS = 1
FPGAs = 18
CALORIMETRY
MUONS
Global Trigger
TMUX=1
FS = 24
FPGAs = 24
EXTERNAL
TRIGGERS
Trigger
accept/reject
12.5 μs
hls4ml

Silicon Tracker
Electromagnetic Calorimeter
Hadron Calorimeter
Superconducting Solenoid
Iron return yoke interspersed with Muon chambers
4T
2T
63 Tb/s
CALORIMETRY (370 FPGAs)
CHARGED PARTICLE TRACKING (174 FPGAs)
MUON CHAMBERS (96 FPGAs)
PARTICLE FLOW (66 FPGAs)
GLOBAL TRIGGER (13 FPGAs)
Accept / Reject
CALORIMETER PRE-PROCESSING
TRACKING PRE-PROCESSING
MUON CHAMBERS PRE-PROCESSING
End-to-end reconstruction model
Latent representation
Downstream Task
Downstream Task
Downstream Task
Downstream Task
12.5 μs
Accept / Reject
Current HL-LHC design
Foundation-model based trigger

Why FPGAs?

Why FPGAs?

- Latency (resource parallelism)

Why FPGAs?

- Throughput (pipeline parallelism)

pipeline

parallelism

Latency, latency, latency (cannot do much on a GPU IN 4 μs)

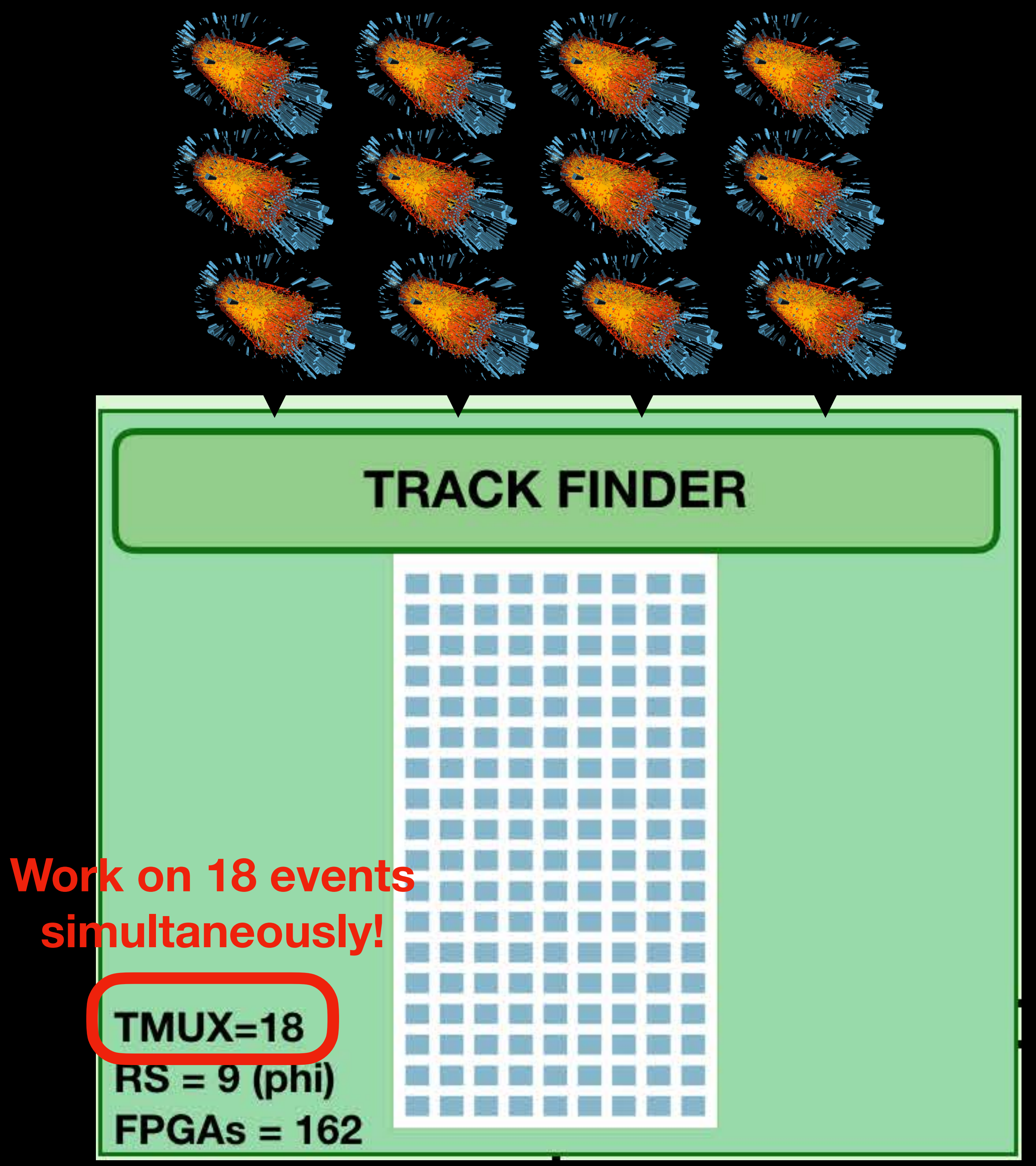
- Can work on different parts of problem, different data simultaneously
- Latency strictly limited by detector frontend buffer

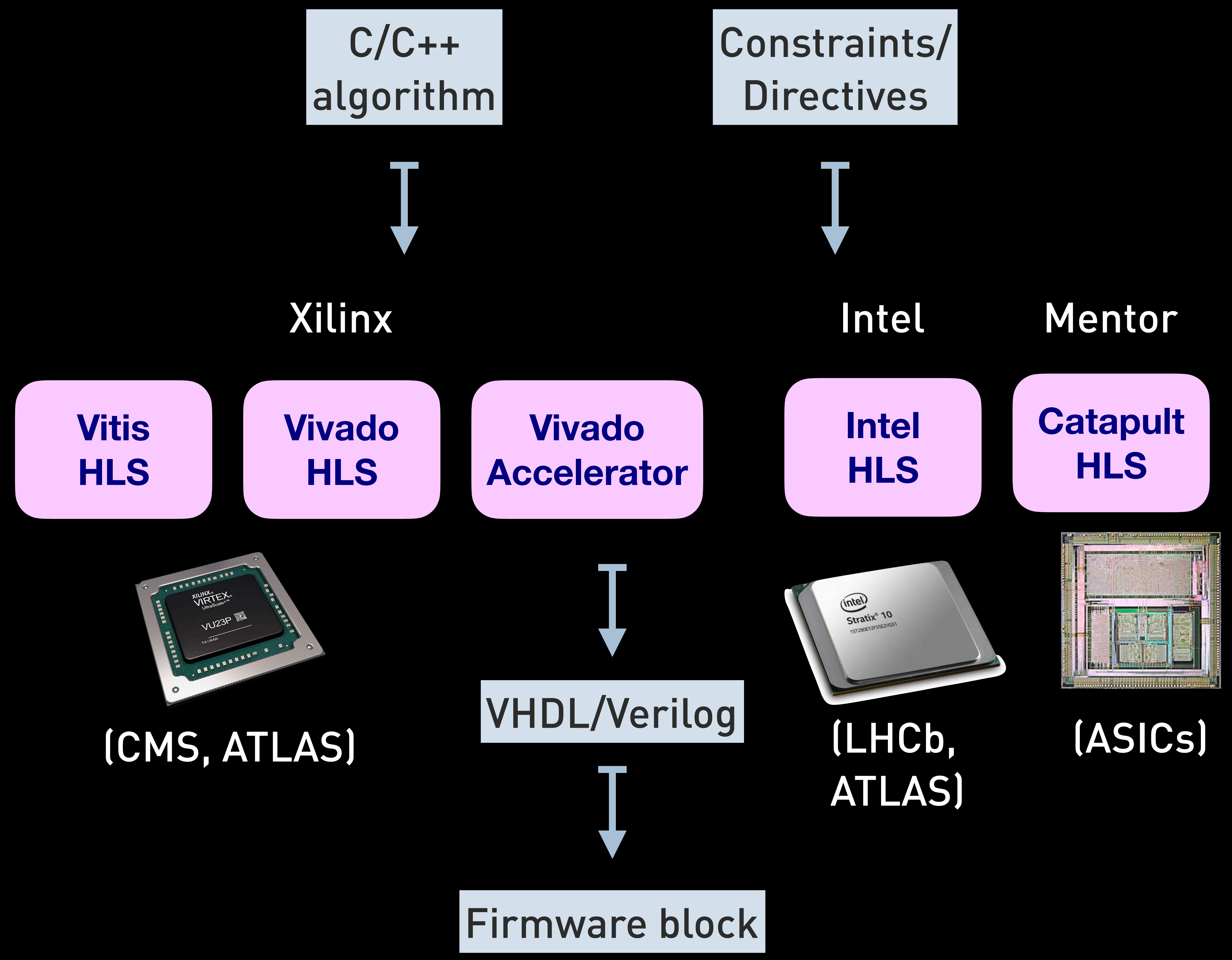
Latency deterministic

- CPU/GPU processing randomness, FPGAs repeatable predictable latency

High bandwidth

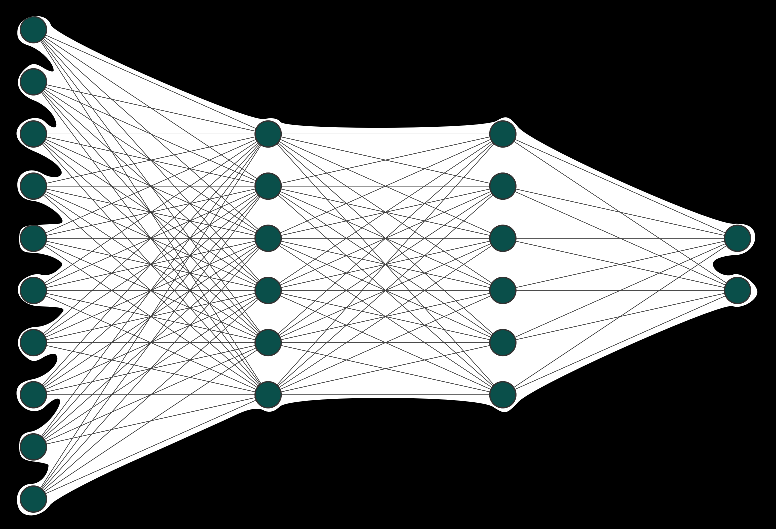
- L1T processes 5% of total internet traffic, dissipate heat of ~7W/cm^2



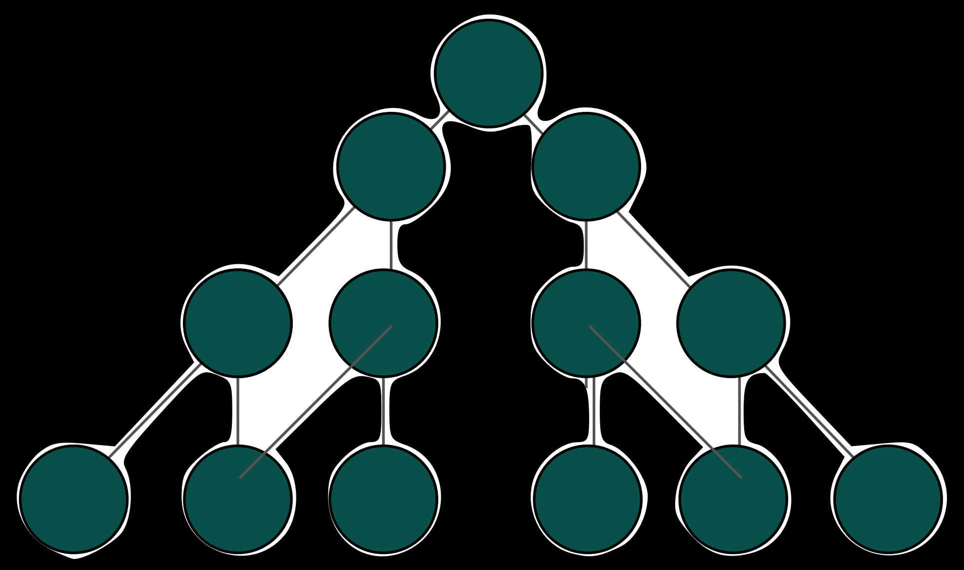

C/C++ algorithm
Constraints/ Directives
Xilinx
Intel
Mentor
Vitis HLS
Vivado HLS
Vivado Accelerator
Intel HLS
Catapult HLS
VIRTEX
VU23P
Stratix 10
(CMS, ATLAS)
VHDL/Verilog
(LHCb, ATLAS)
(ASICs)
Firmware block

https://github.com/fastmachinelearning/hls4ml
https://fastmachinelearning.org/hls4ml/

KERAS / PyTorch / ONNX



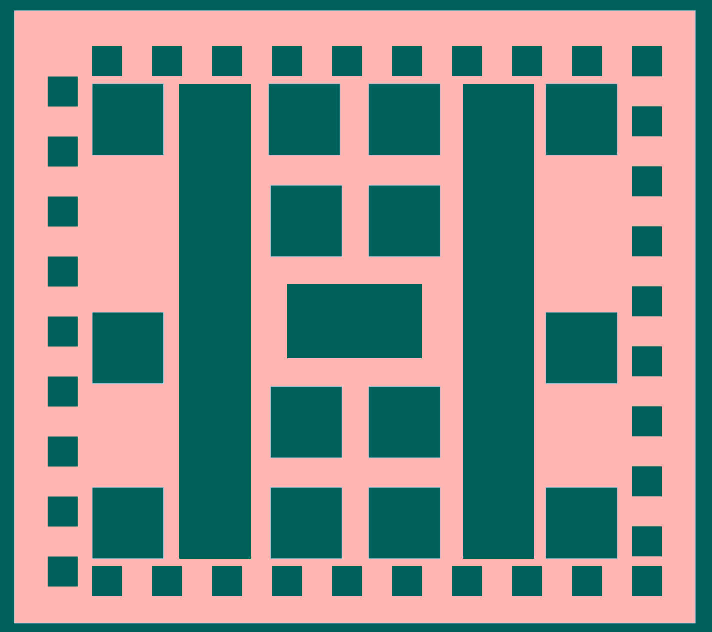
TensorFlow DF / scikit-learn / XGBoost



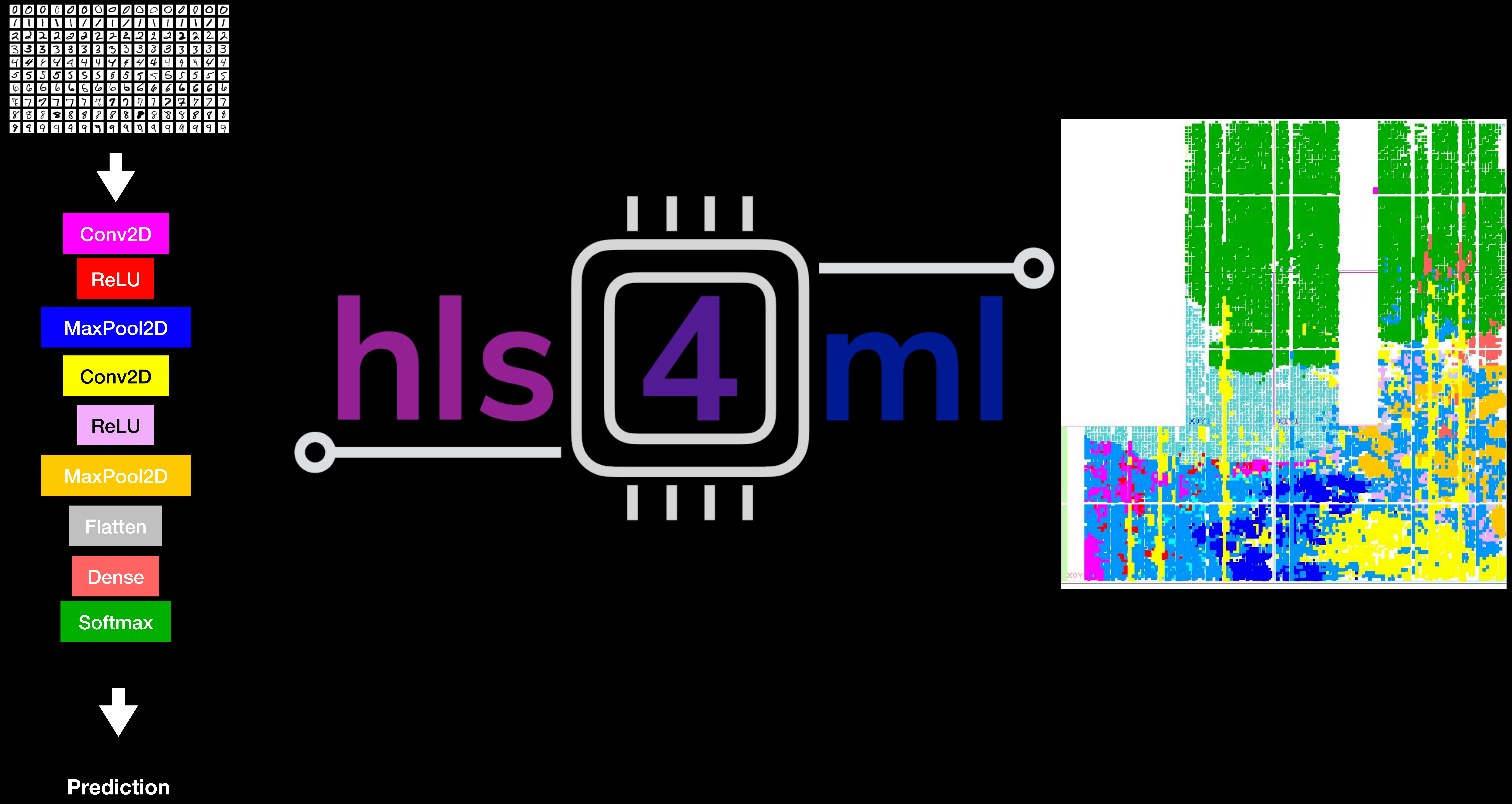
hls4ml

Conifer

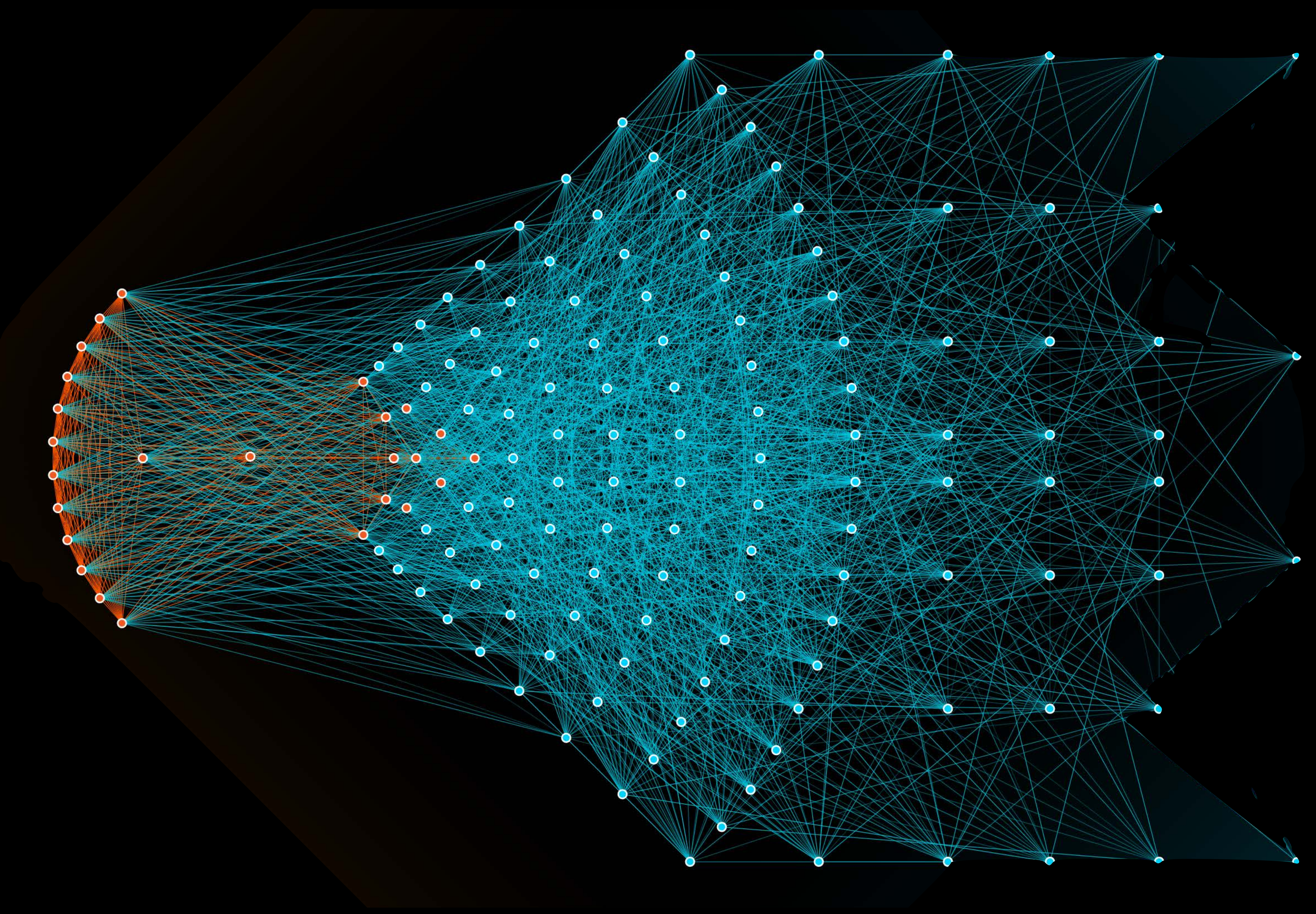
HLS project:
Vivado / Vitis / Intel Quartus / IntelOne API / Catapult



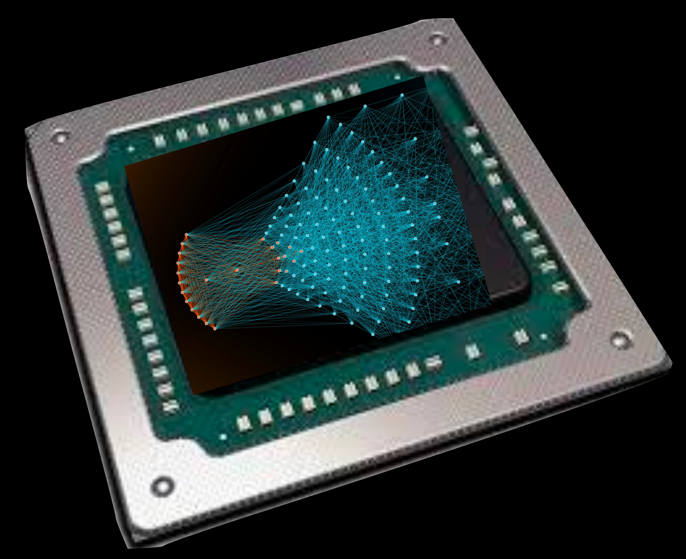

```
pip install hls4ml
pip install conifer
```



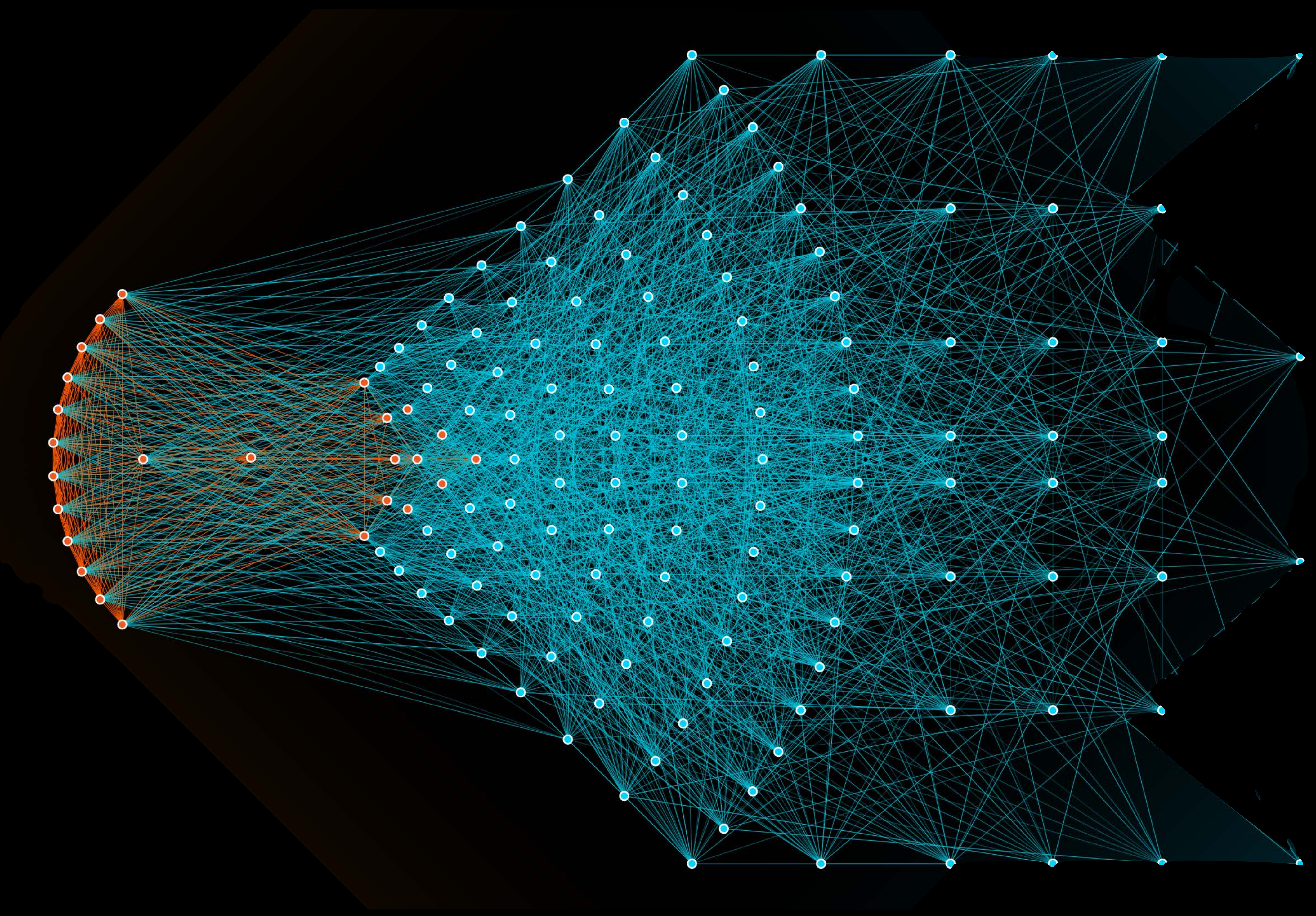
Conv2D
ReLU
MaxPool2D
Conv2D
ReLU
MaxPool2D
Flatten
Dense
Softmax
Prediction
hls4ml



Ideally

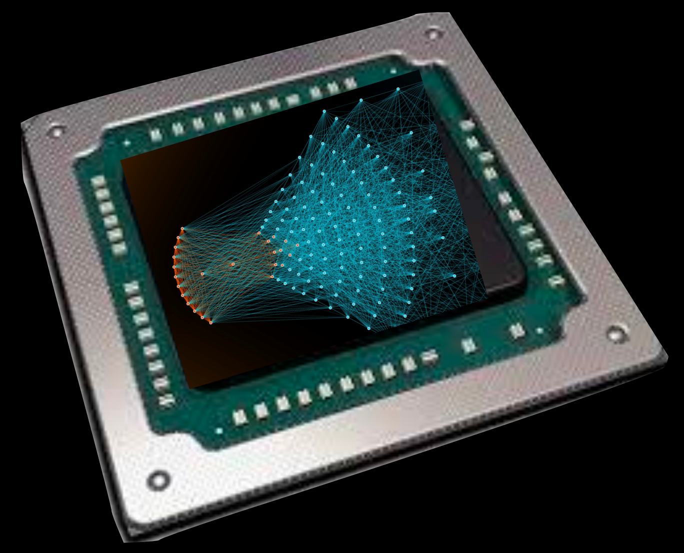


Reality



Ideally

- Quantization
- Pruning
- Parallelisation
- Knowledge distillation

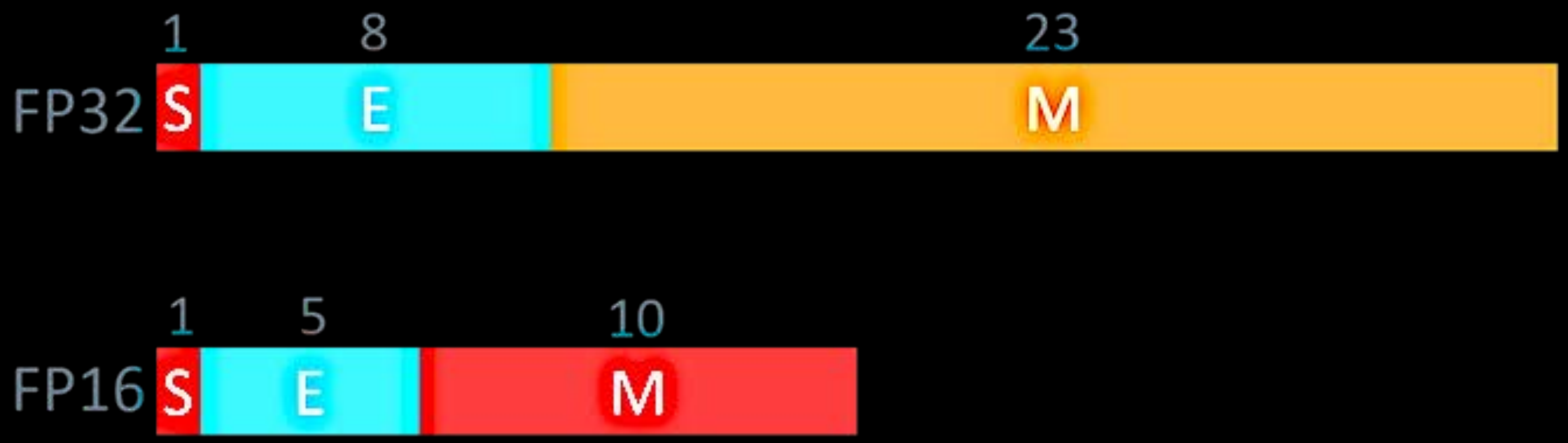


Reality

GPT-3

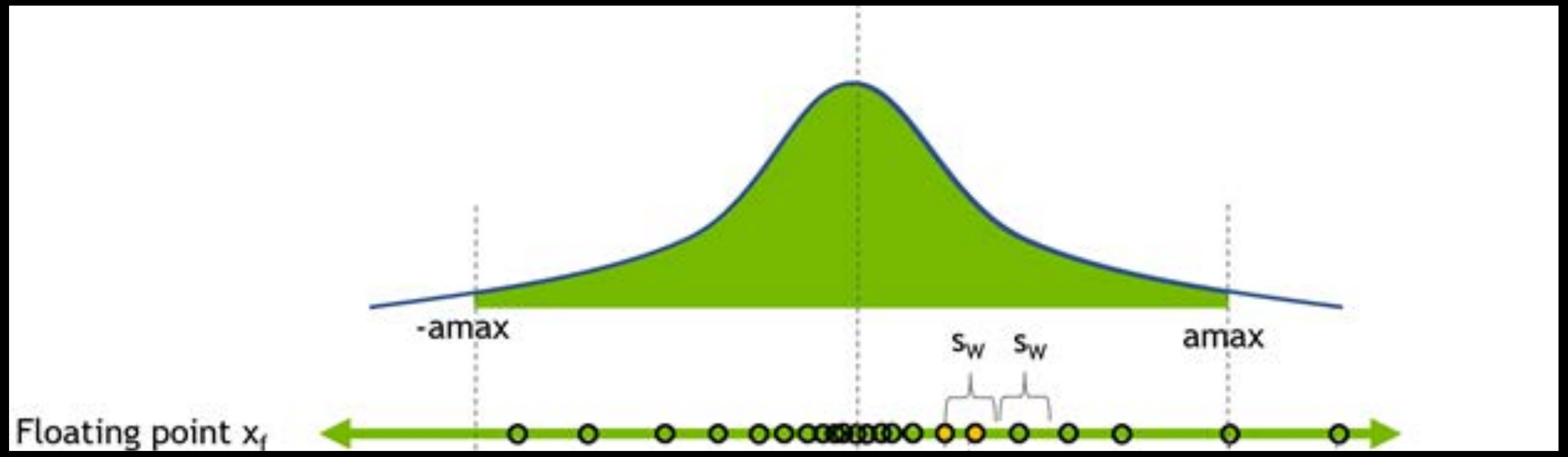
175,000,000,000

FP16 vs FP32



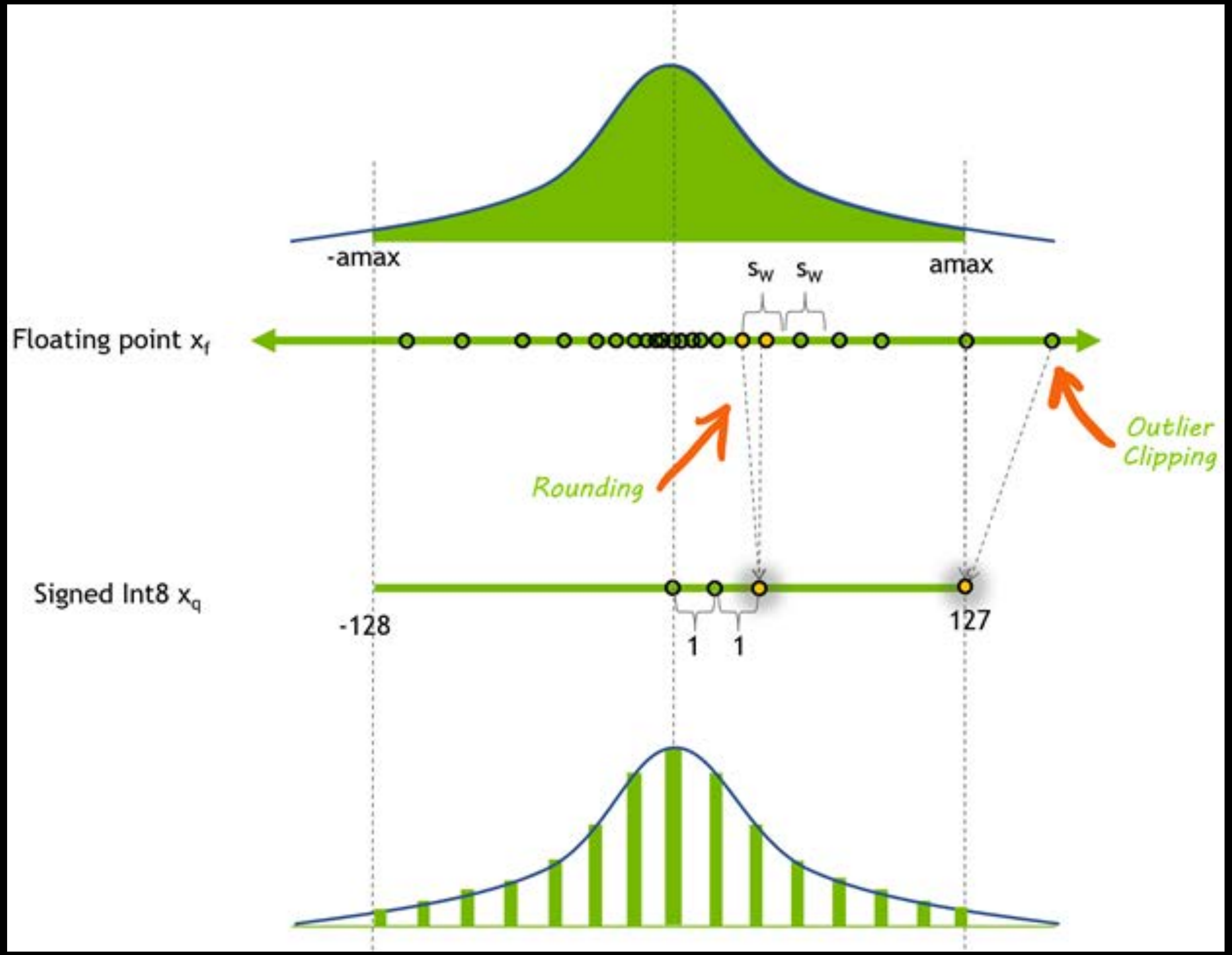
	Range	Accuracy		
FP32	10^{-38} - 10^{38}	0.000006%	→	~700 GB
FP16	$6x10^{-5}$ - $6x10^{4}$	0.05%	→	~350 GB

Quantization



Floating point 32:
4B numbers in [-3.4e38, +3.4e38]

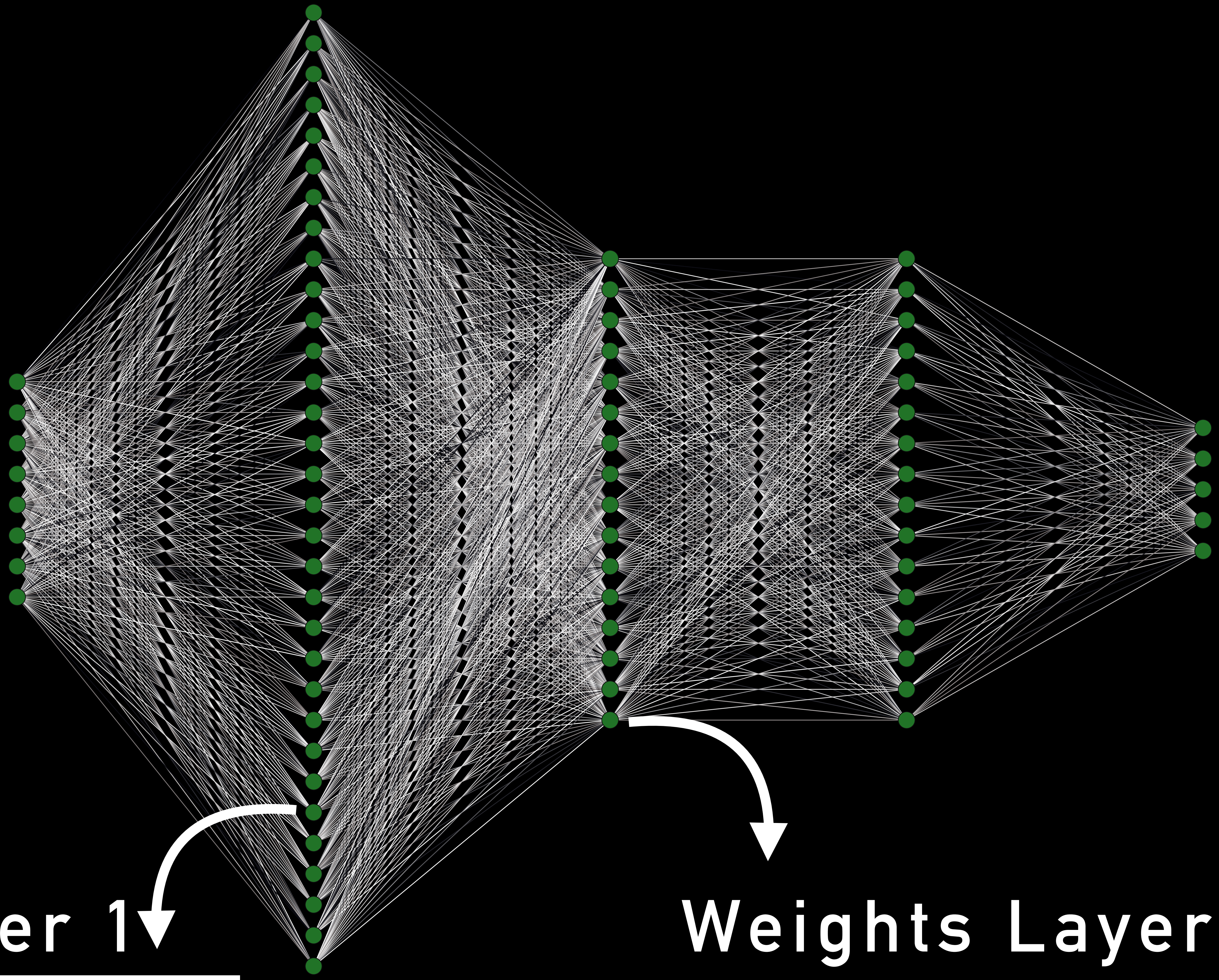
Quantization



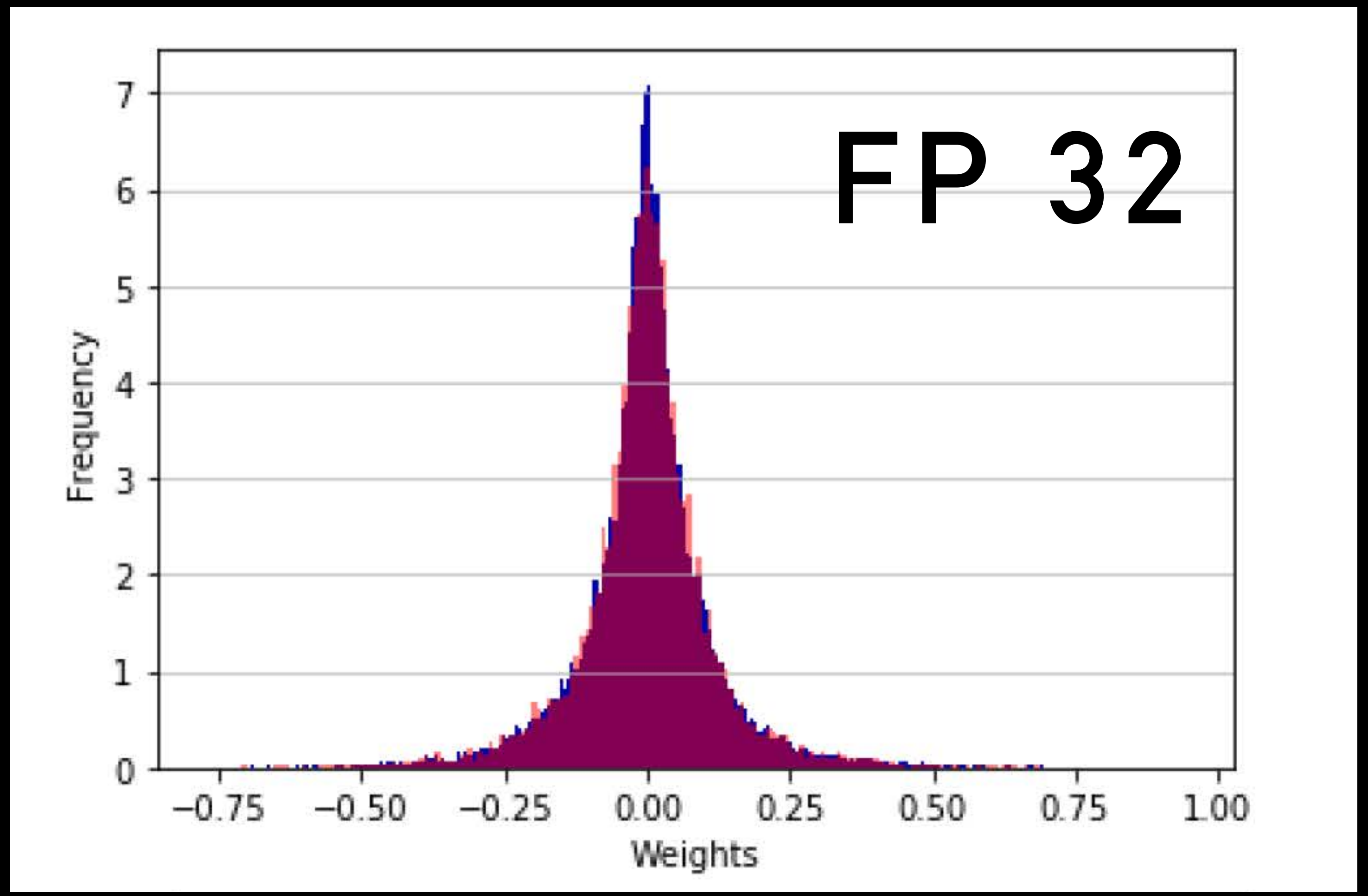
Quantising:

int8 2^8=256 numbers in [-128,127]

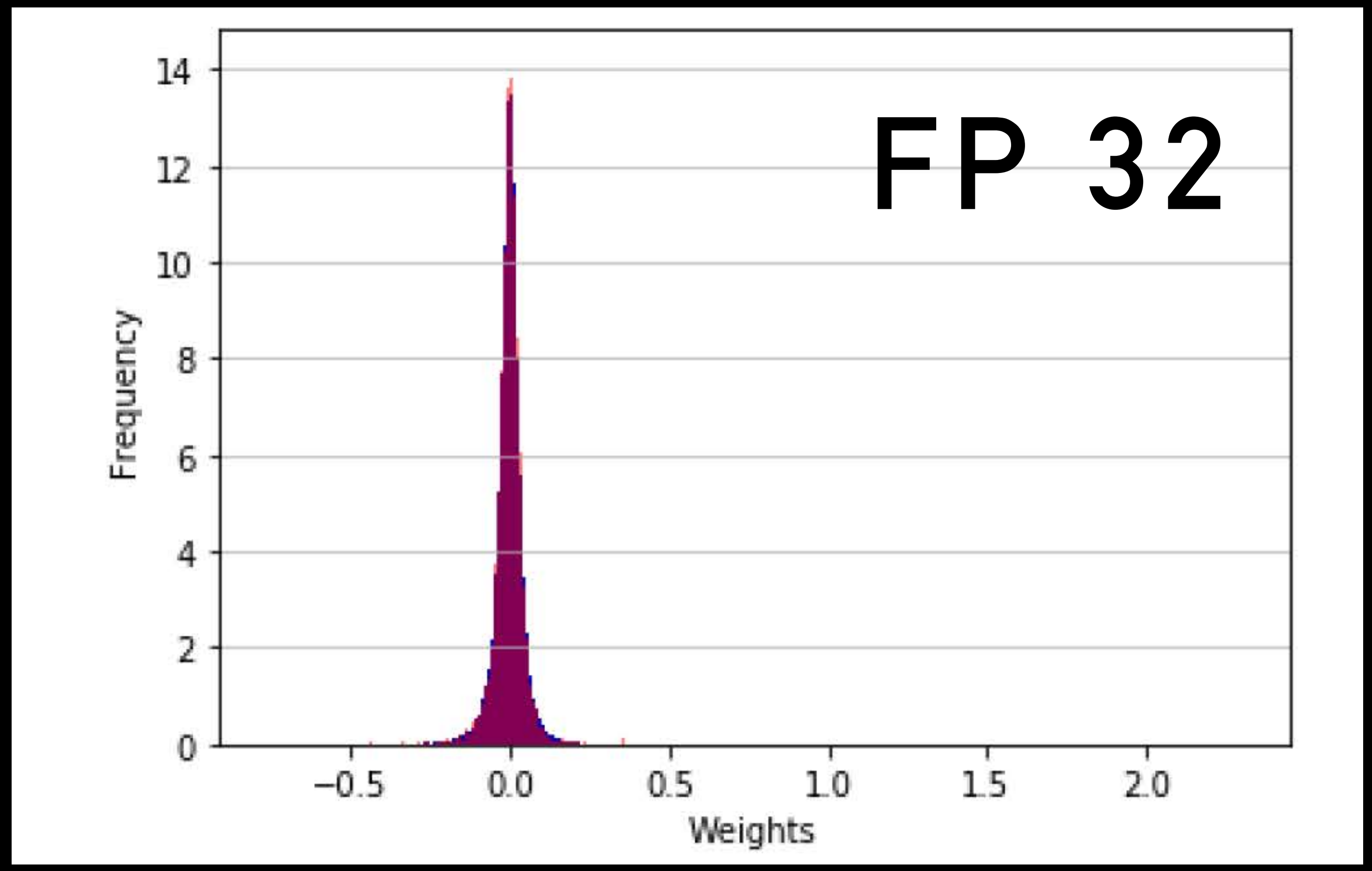
$$x_q = Clip(Round(\frac{x_f}{scale}))$$



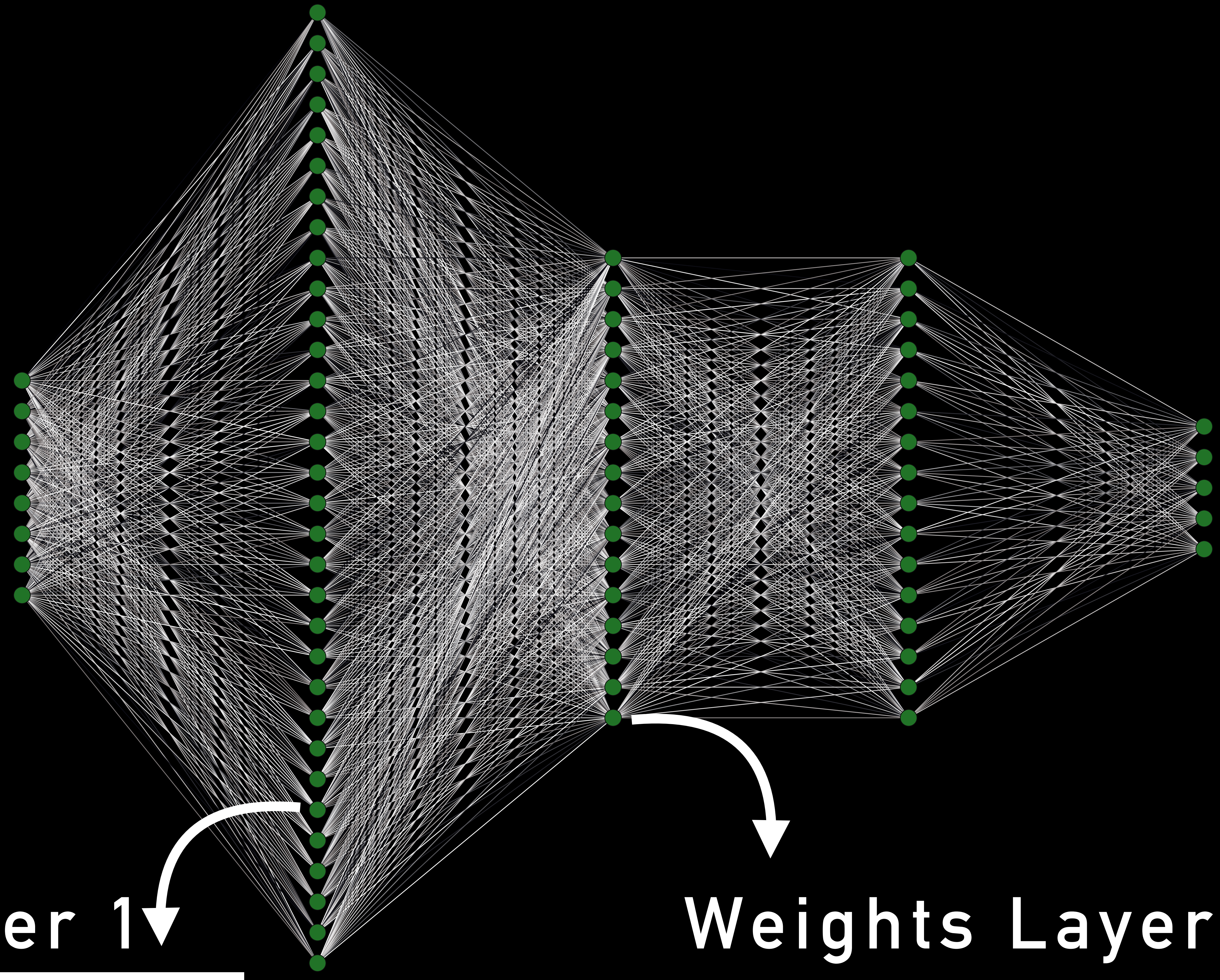
Weights Layer 1



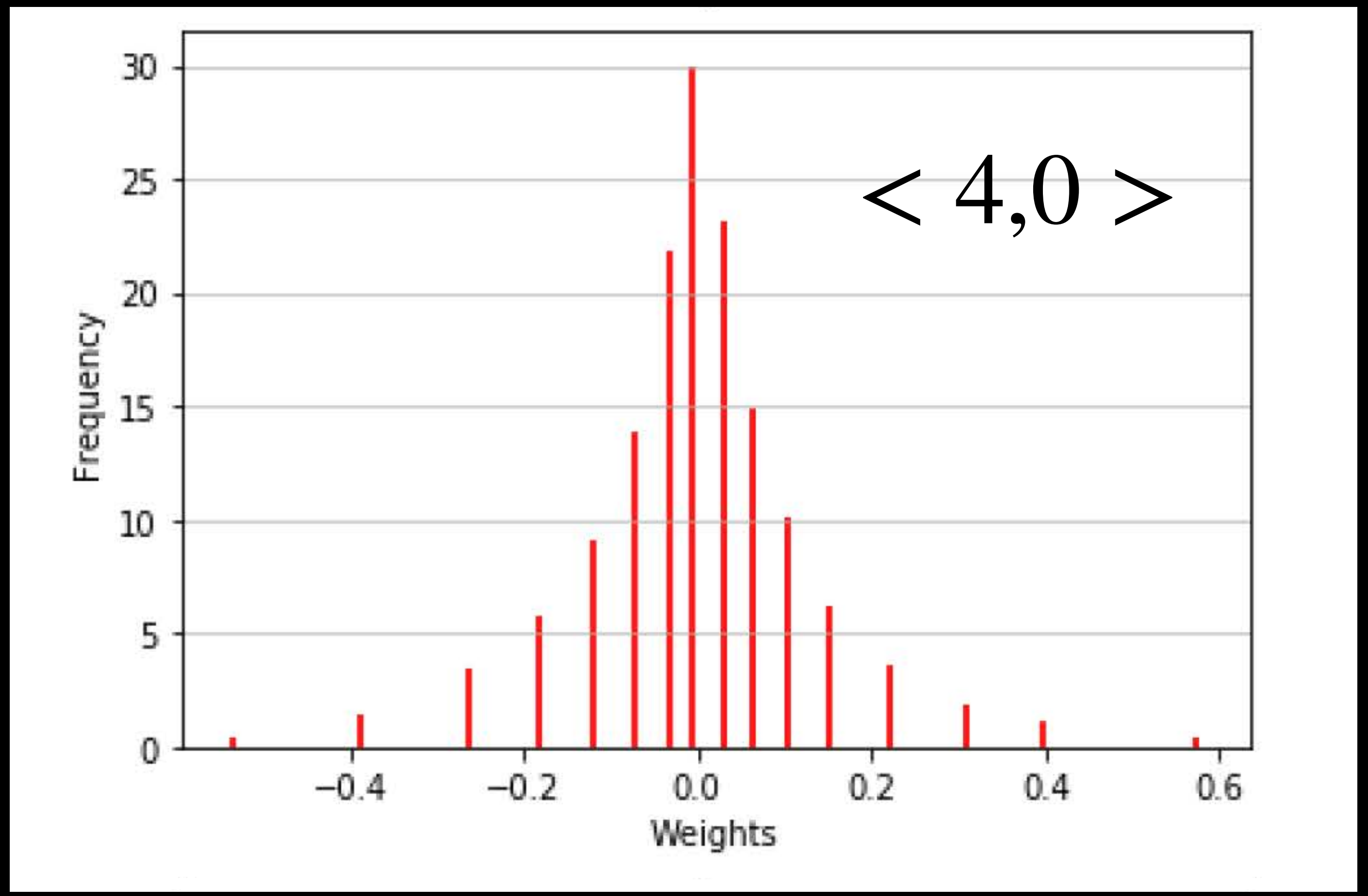
Weights Layer 2



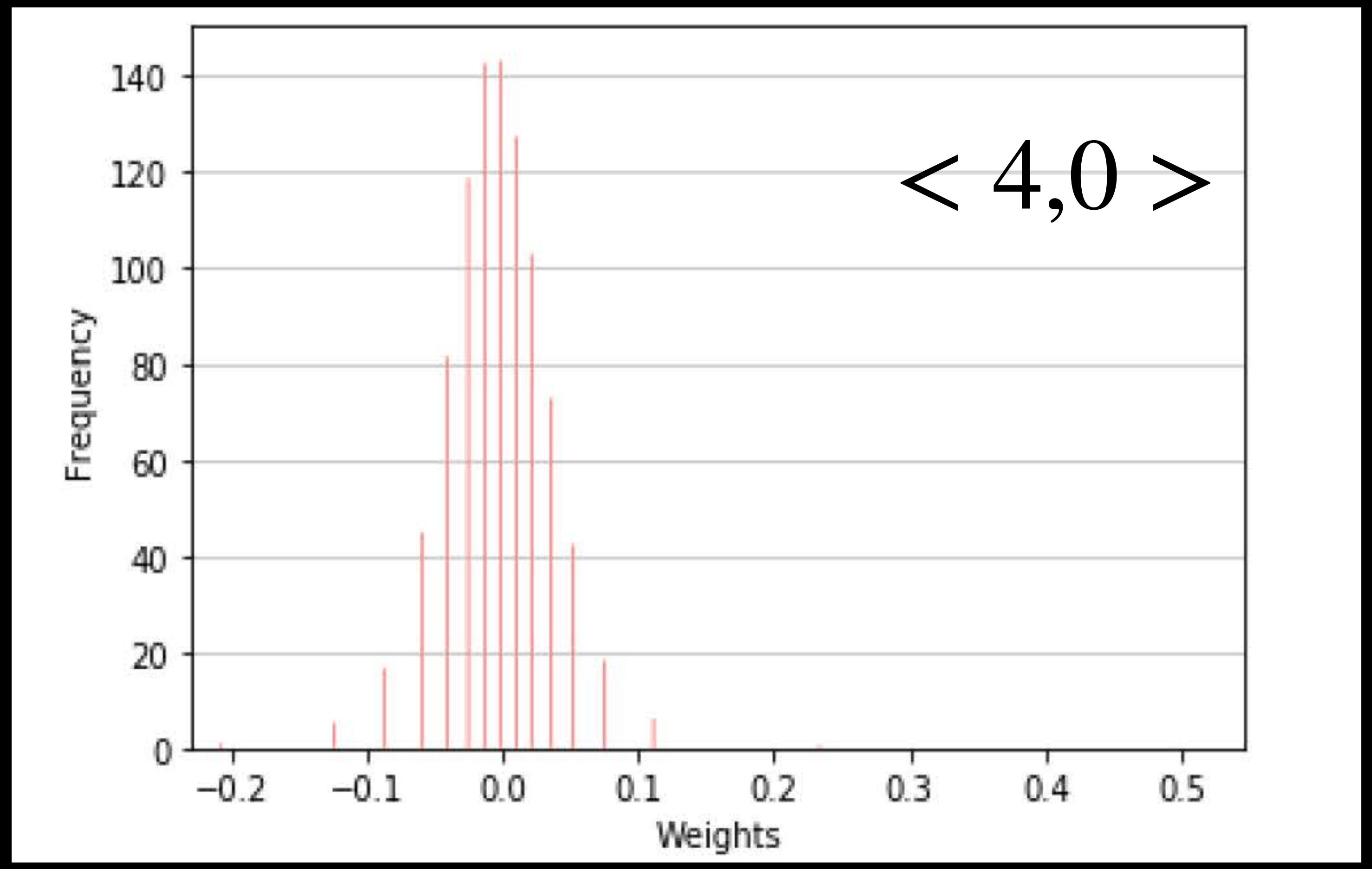
Fixed point



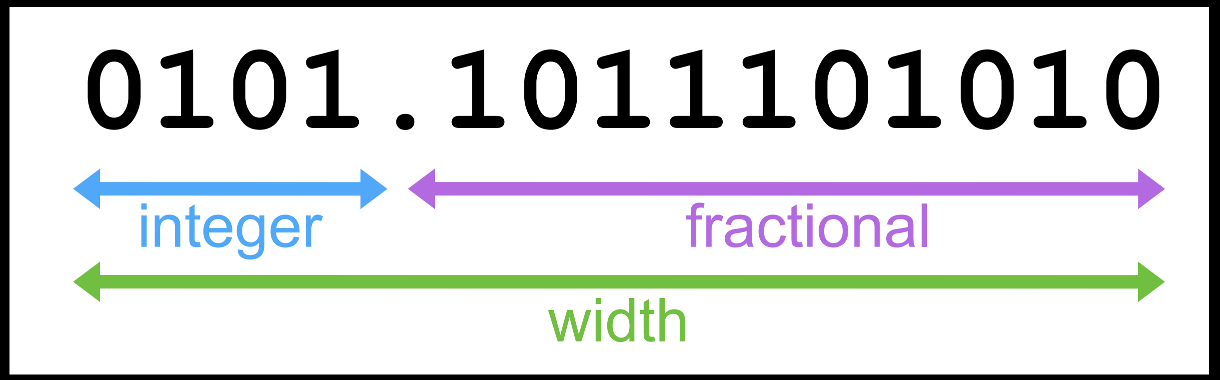
Weights Layer 1

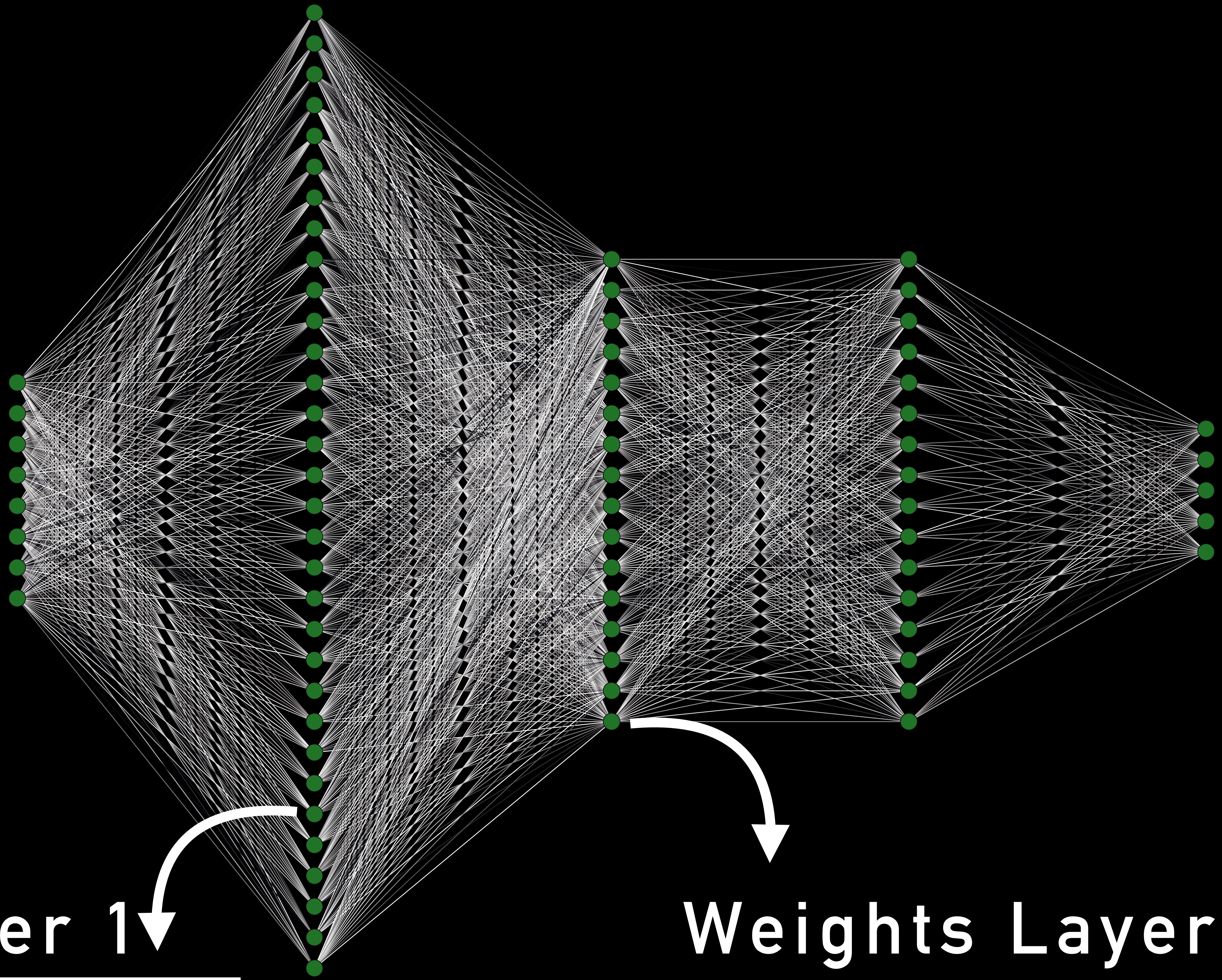


Weights Layer 2

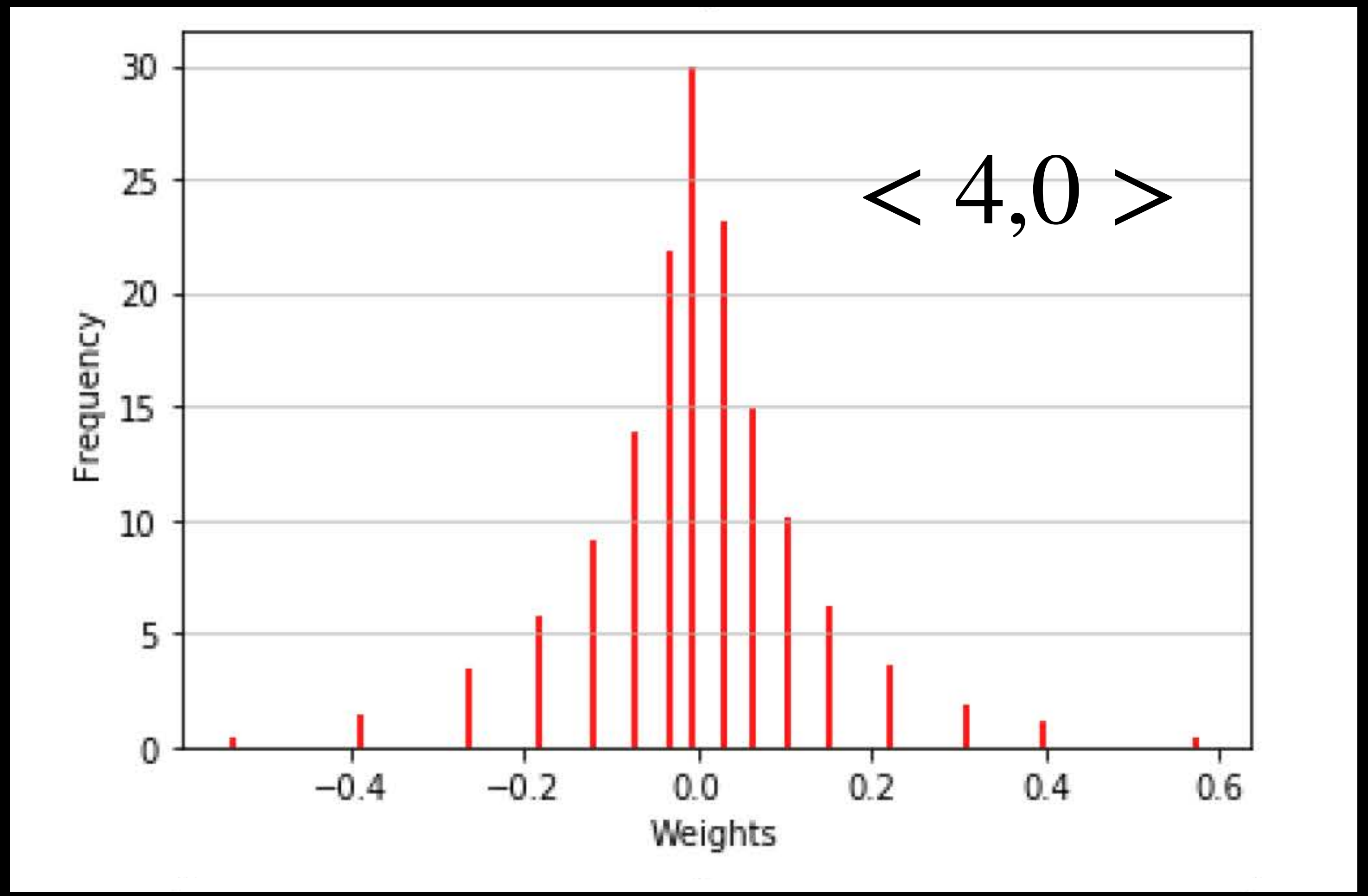


Fixed point

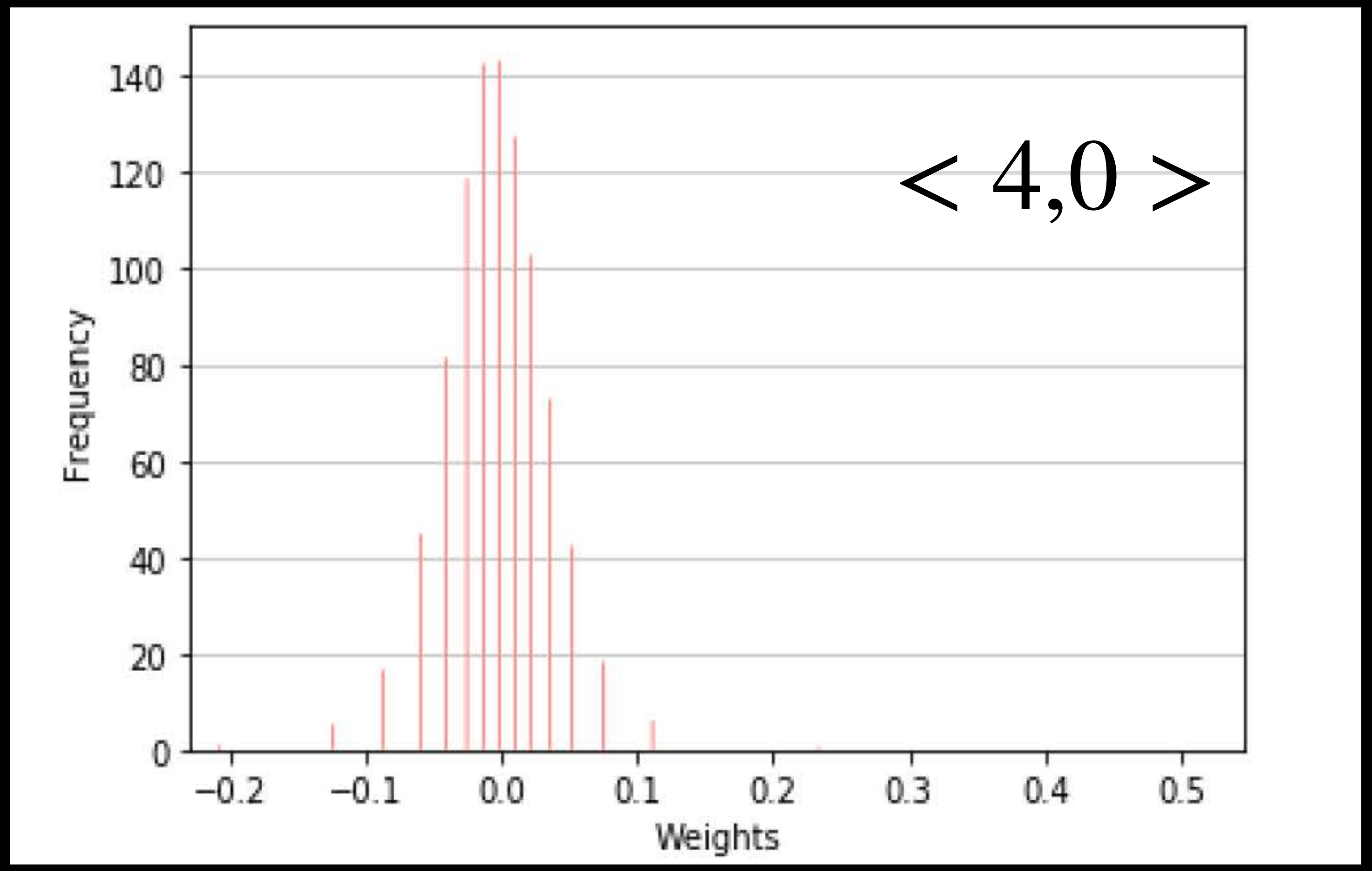




Weights Layer 1



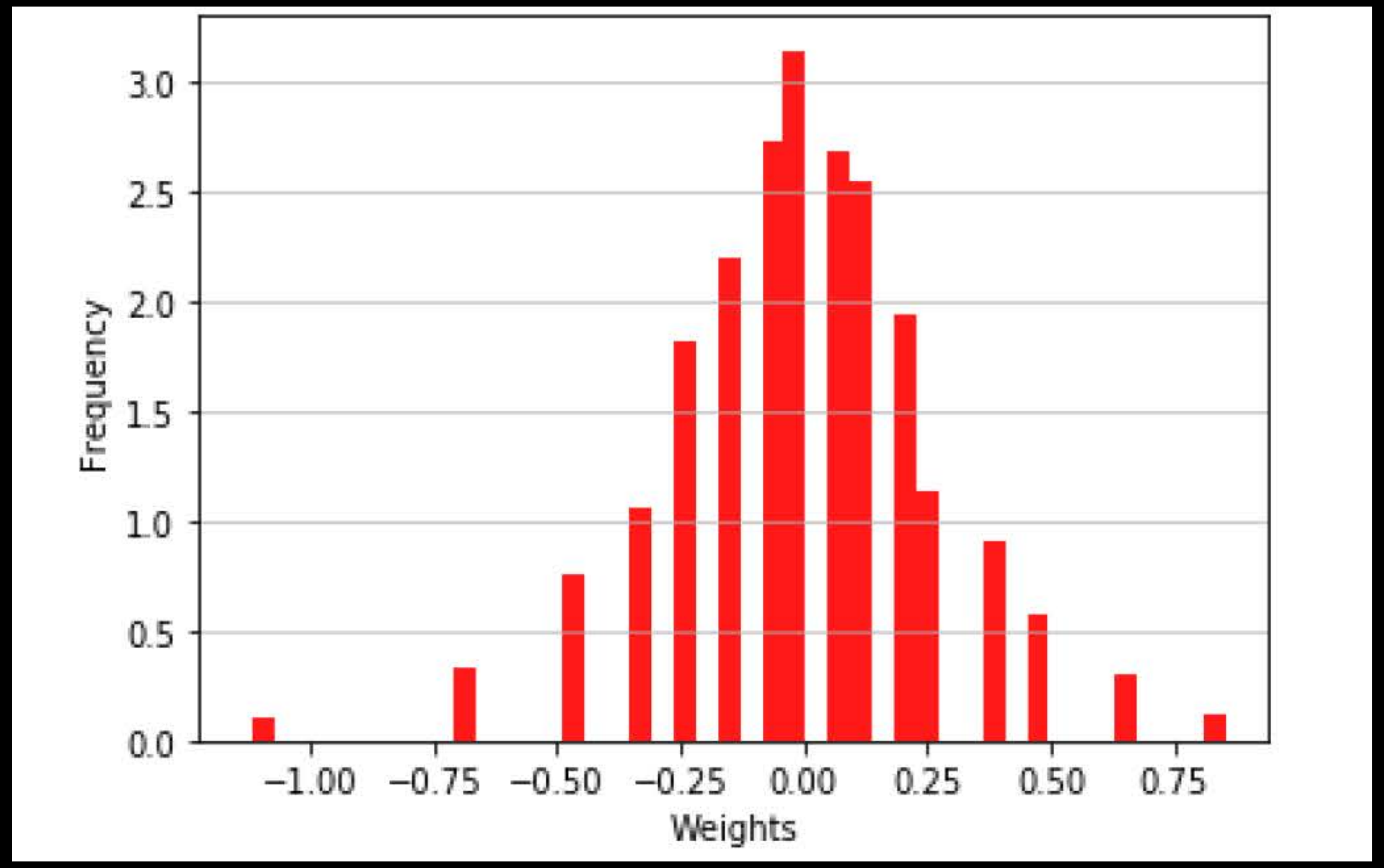
Weights Layer 2

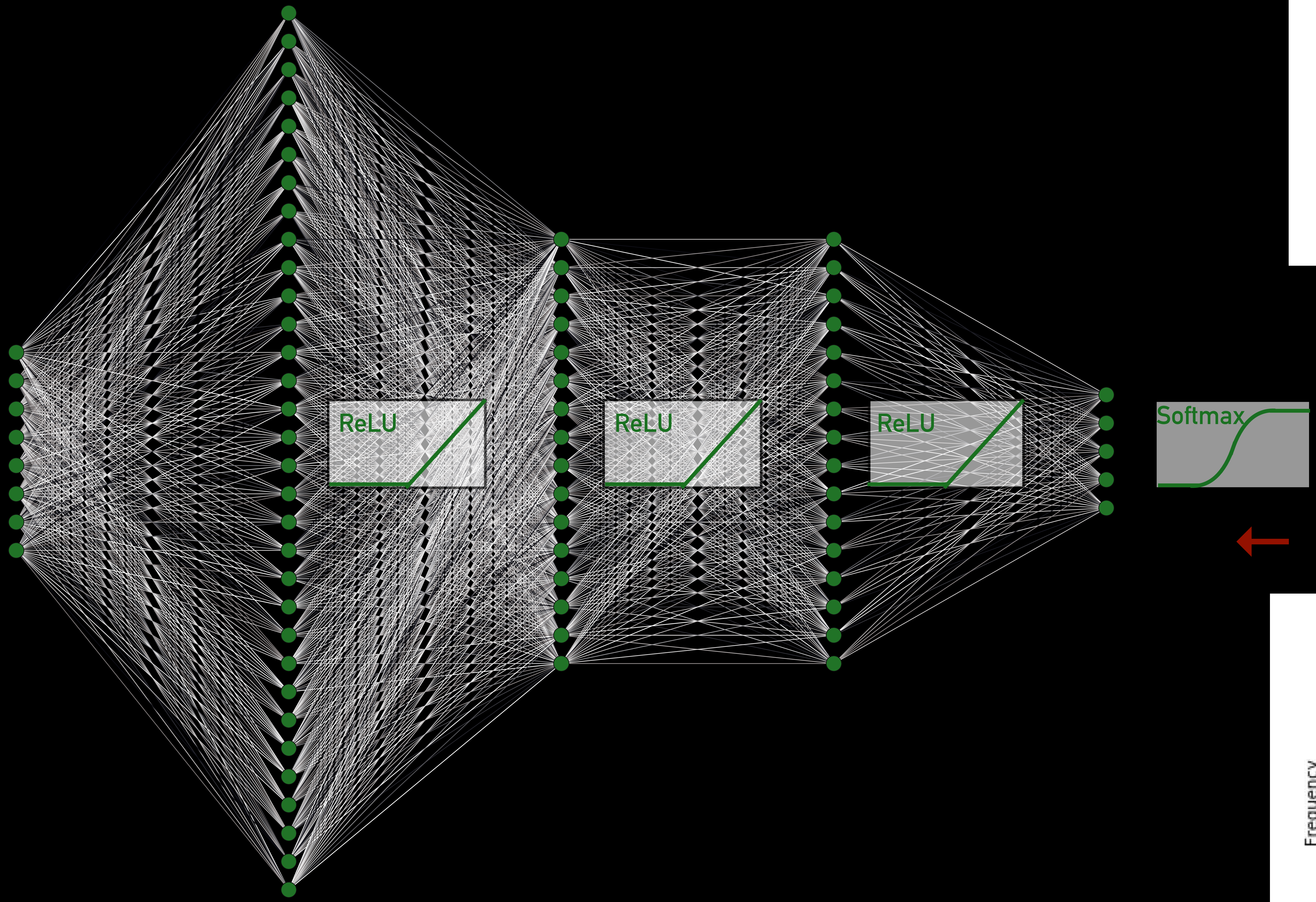


hls4ml + Google

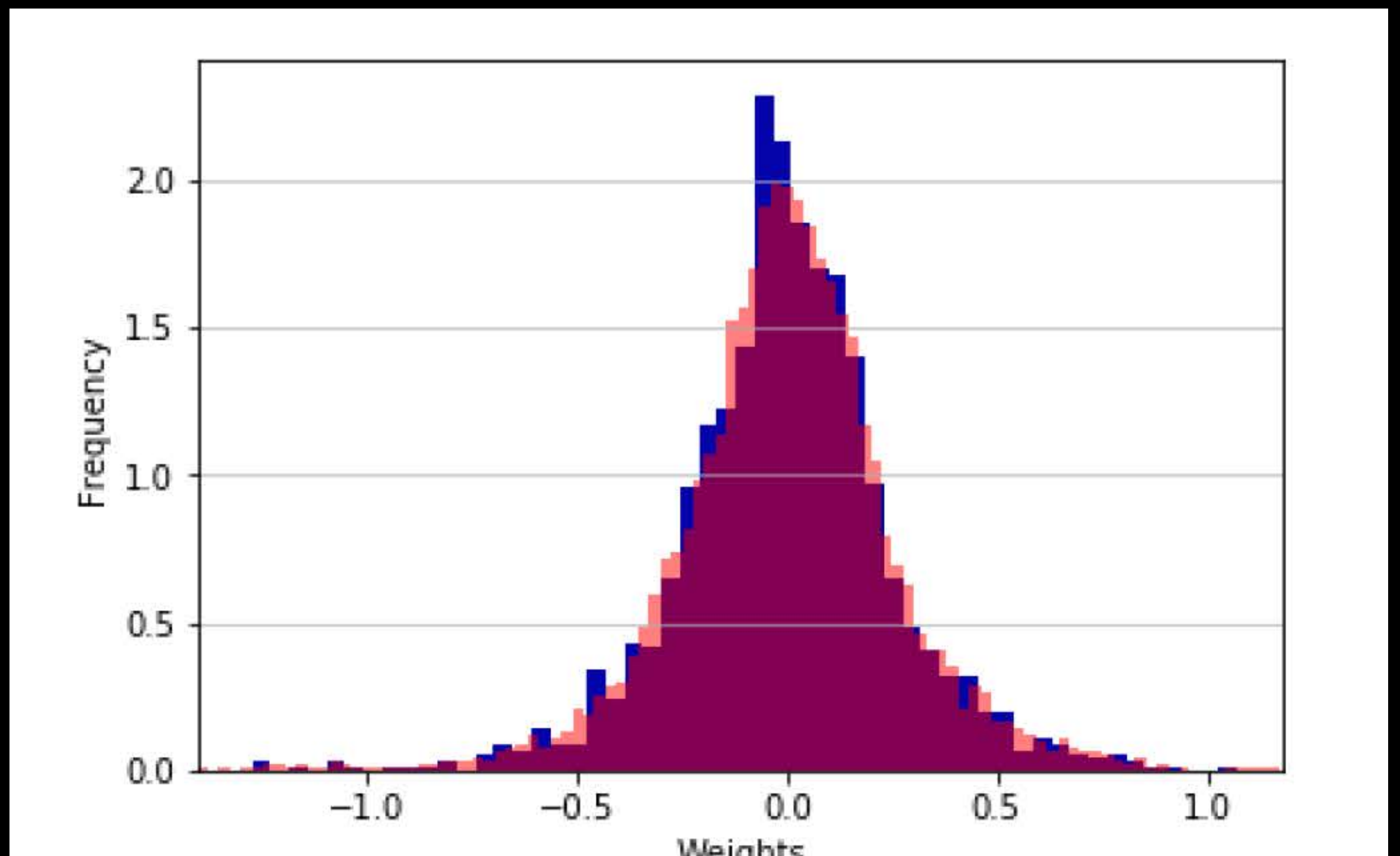
Quantization-aware training

Forward pass →





← Back propagation



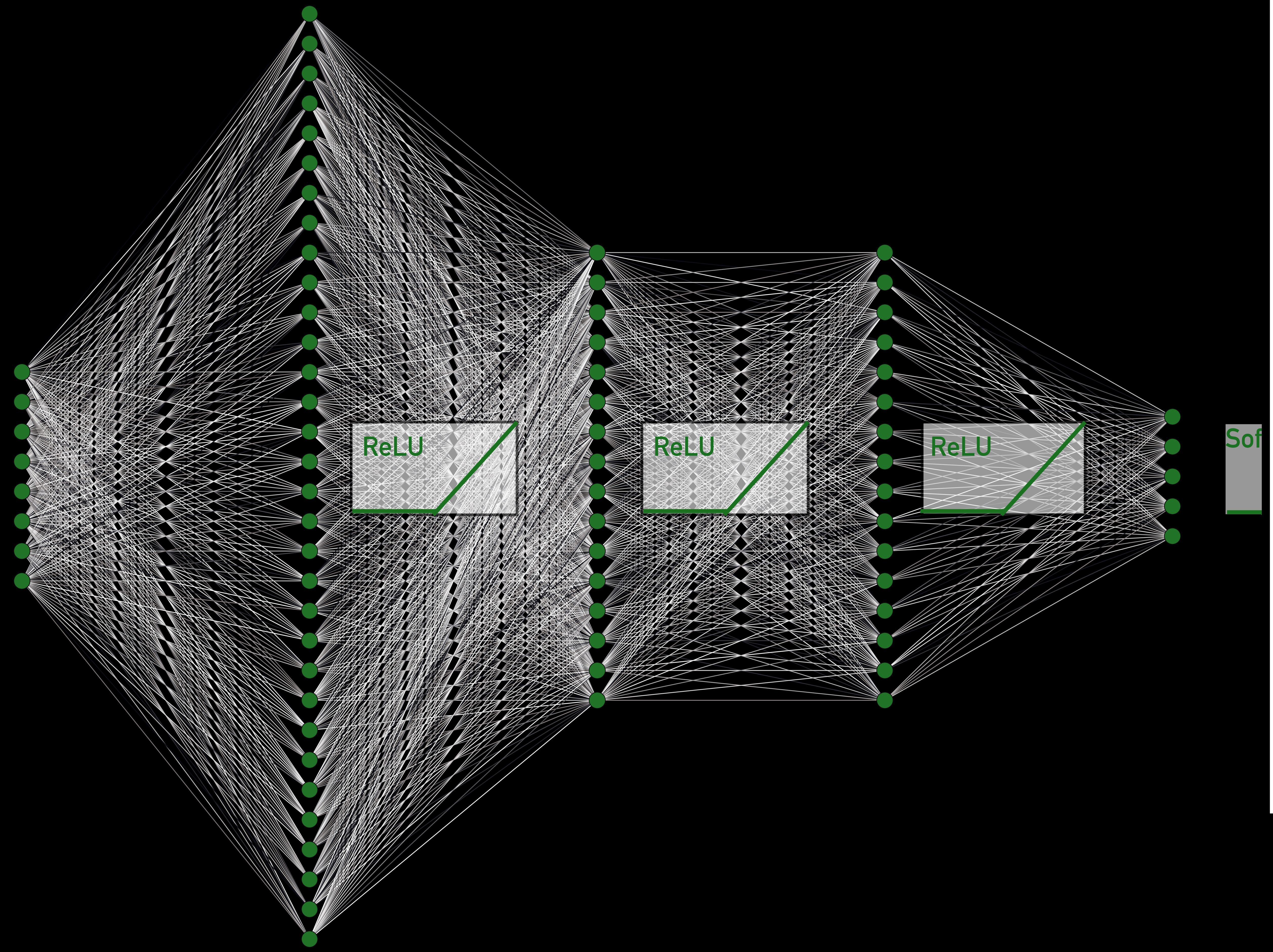
Nature Machine Intelligence 3 (2021)

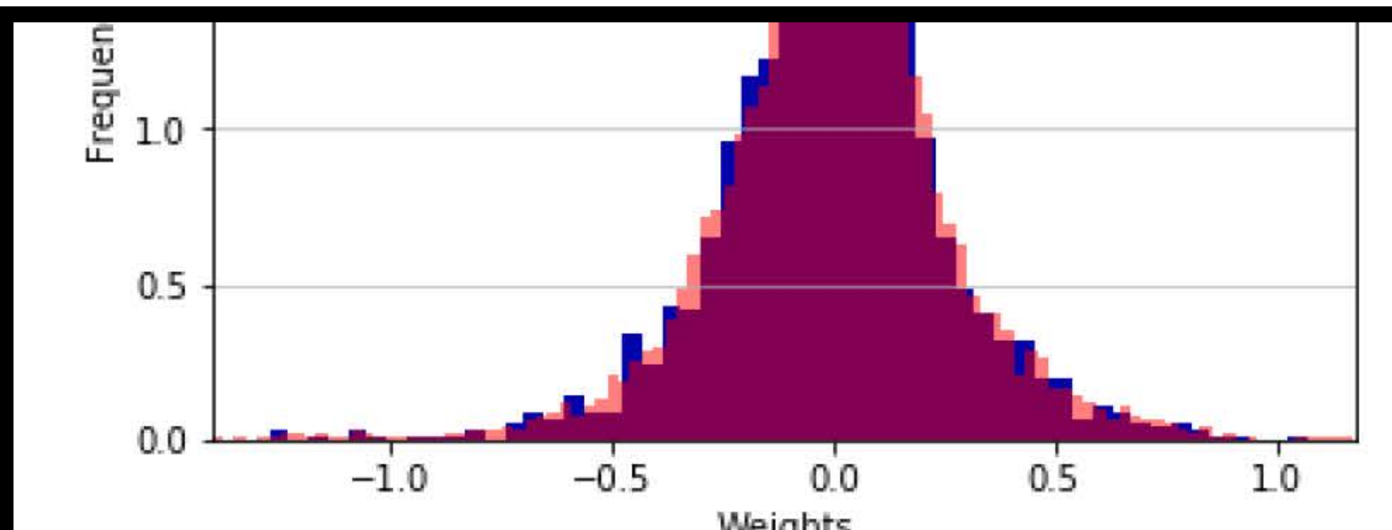
hls4ml + Google

Quantization-aware training

Forward pass →

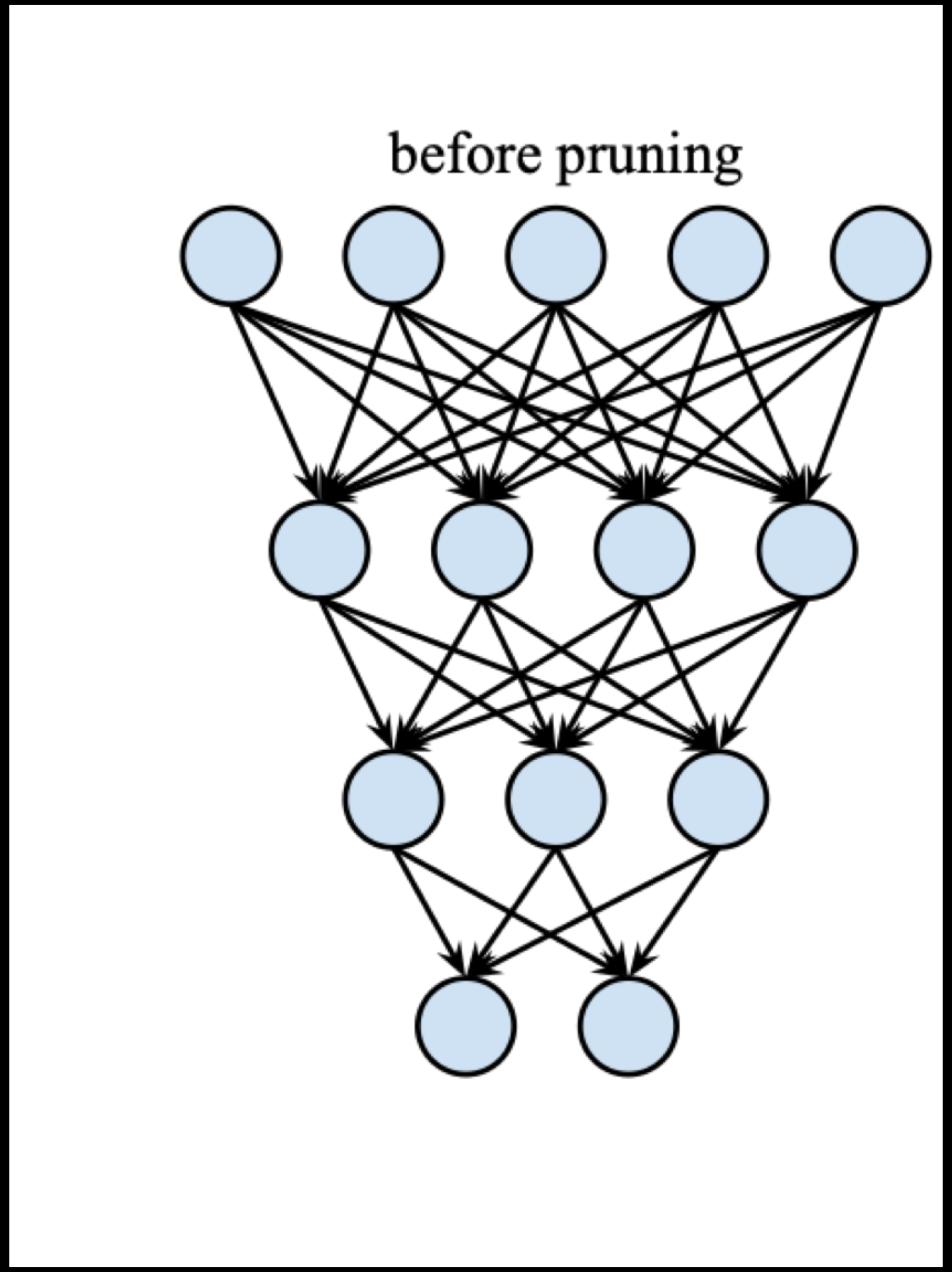

```
from tensorflow.keras.layers import Input, Activation
from qkeras import quantized_bits
from qkeras import QDense, QActivation
from qkeras import QBatchNormalization

x = Input((16))
x = QDense(64,
    kernel_quantizer = quantized_bits(6,0,alpha=1),
    bias_quantizer   = quantized_bits(6,0,alpha=1))(x)
x = QBatchNormalization()(x)
x = QActivation('quantized_relu(6,0)')(x)
x = QDense(32,
    kernel_quantizer = quantized_bits(6,0,alpha=1),
    bias_quantizer   = quantized_bits(6,0,alpha=1))(x)
x = QBatchNormalization()(x)
x = QActivation('quantized_relu(6,0)')(x)
x = QDense(32,
    kernel_quantizer = quantized_bits(6,0,alpha=1),
    bias_quantizer   = quantized_bits(6,0,alpha=1))(x)
x = QBatchNormalization()(x)
x = QActivation('quantized_relu(6,0)')(x)
x = QDense(5,
    kernel_quantizer = quantized_bits(6,0,alpha=1),
    bias_quantizer   = quantized_bits(6,0,alpha=1))(x)
x = Activation('softmax')(x)
```

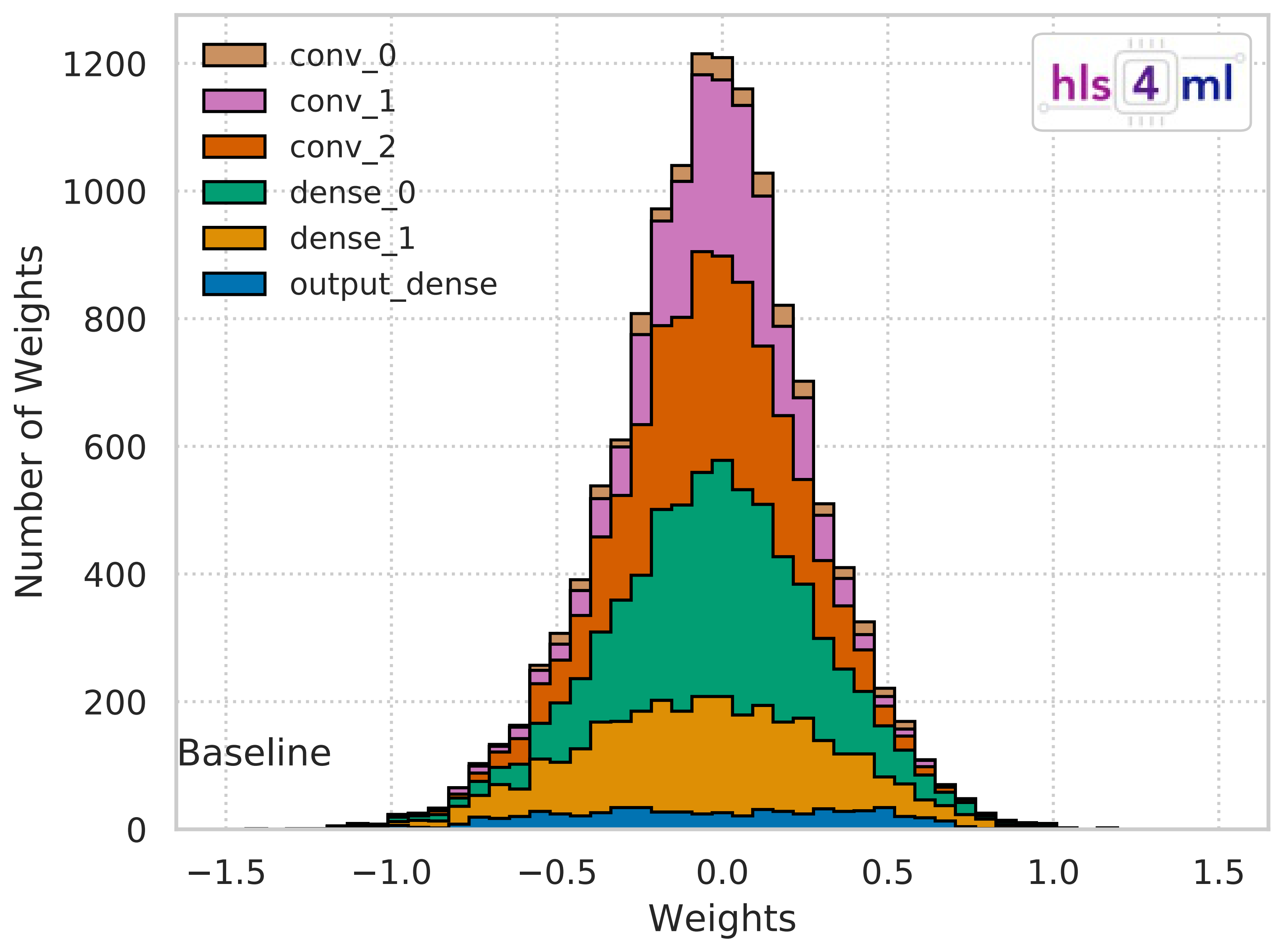




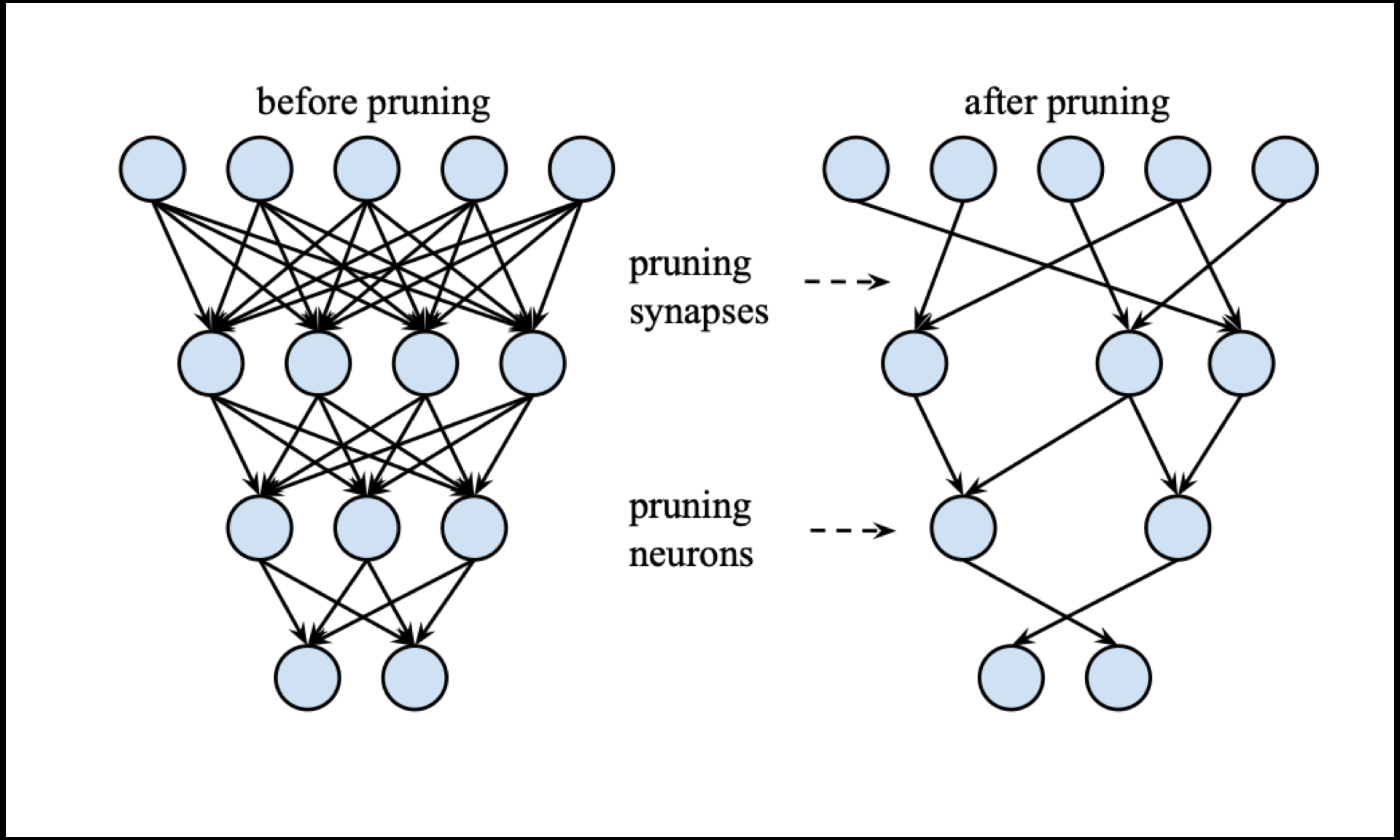
Nature Machine Intelligence 3 (2021)

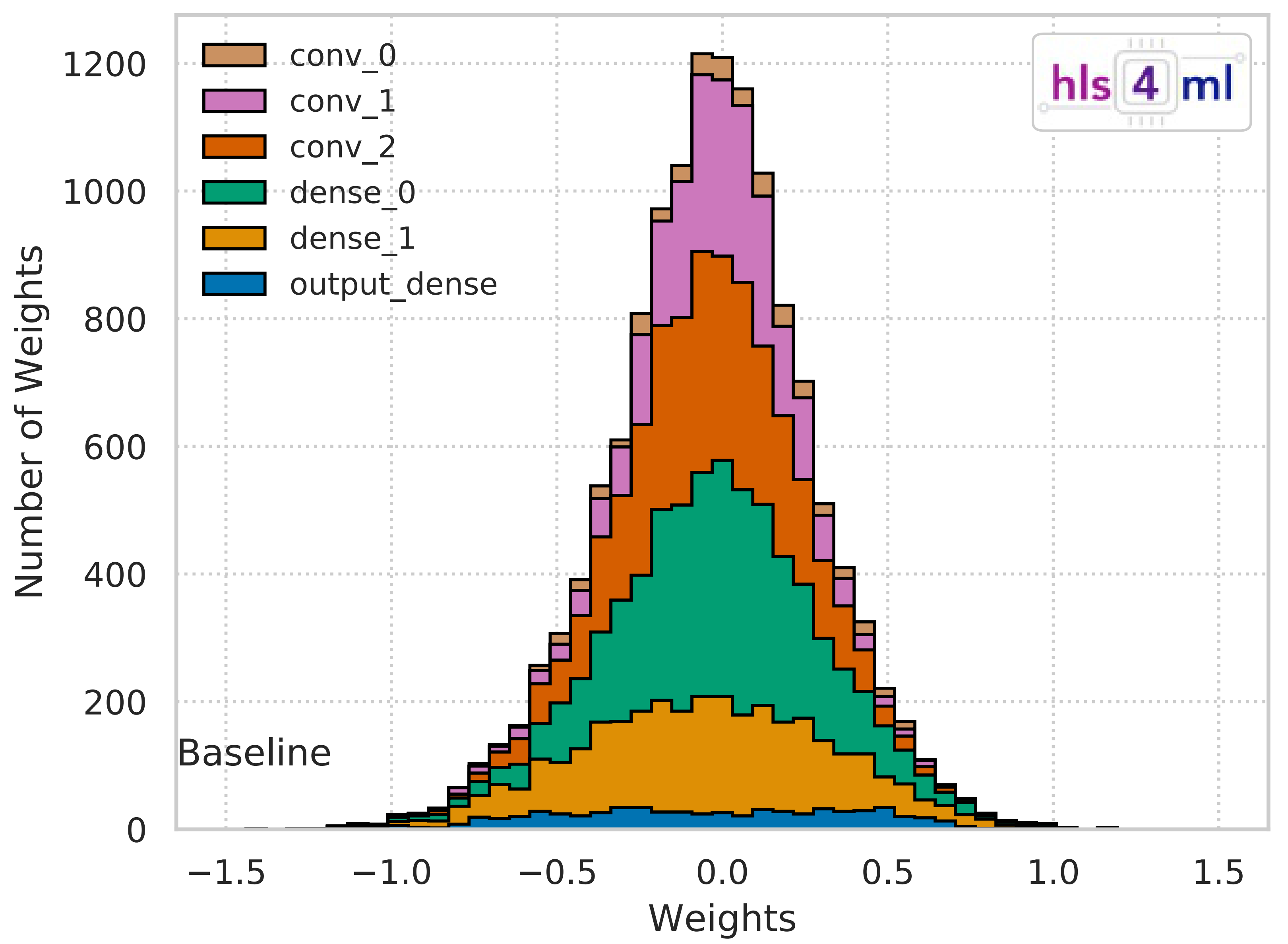
Pruning

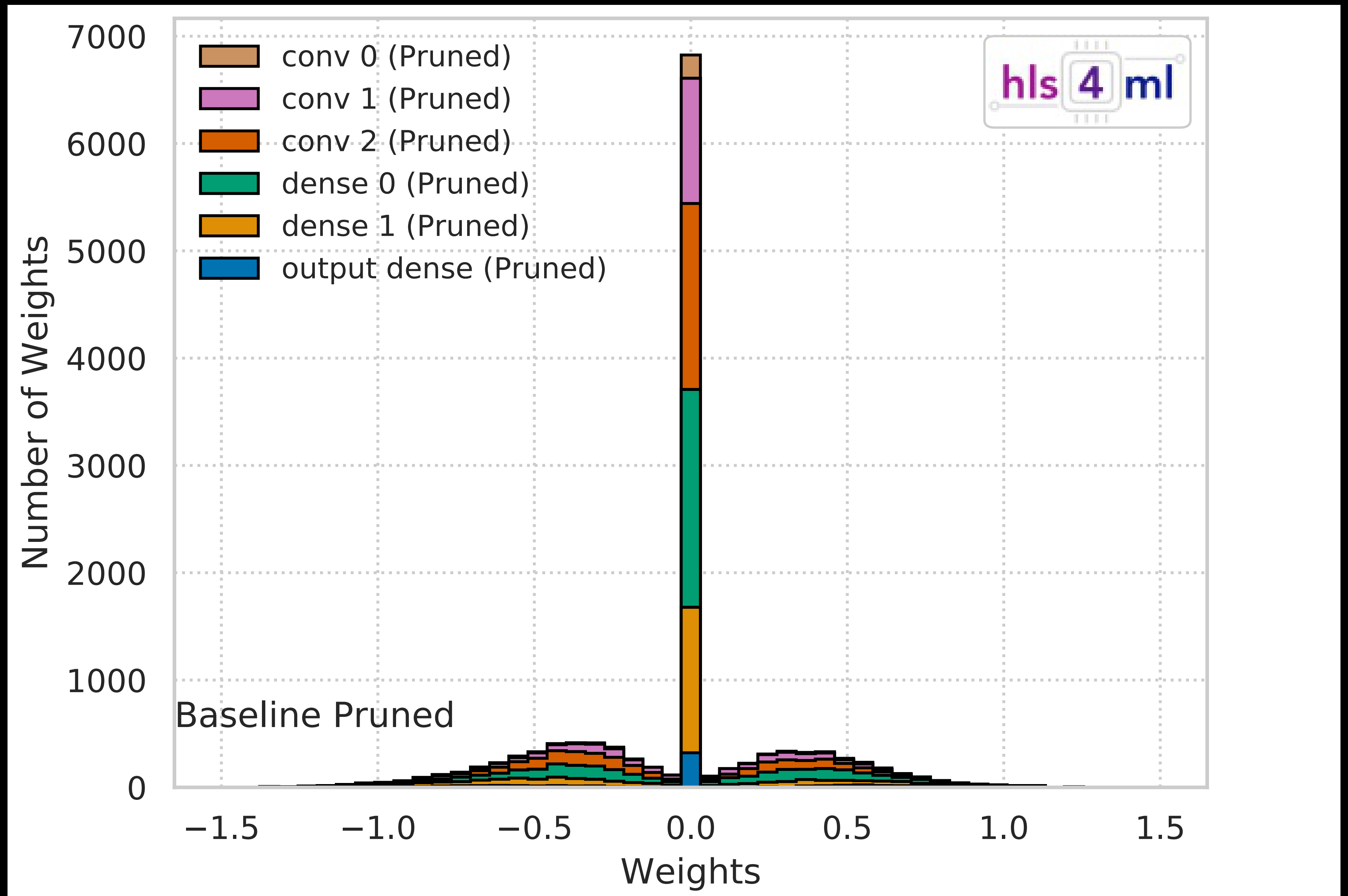

before pruning


conv_0
conv_1
conv_2
dense_0
dense_1
output_dense
hls 4 ml
Baseline
Number of Weights
Weights
0
200
400
600
800
1000
1200
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

Pruning


before pruning
after pruning
pruning
synapses
pruning
neurons


conv_0
conv_1
conv_2
dense_0
dense_1
output_dense
hls 4 ml
Baseline
Number of Weights
Weights
0
200
400
600
800
1000
1200
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

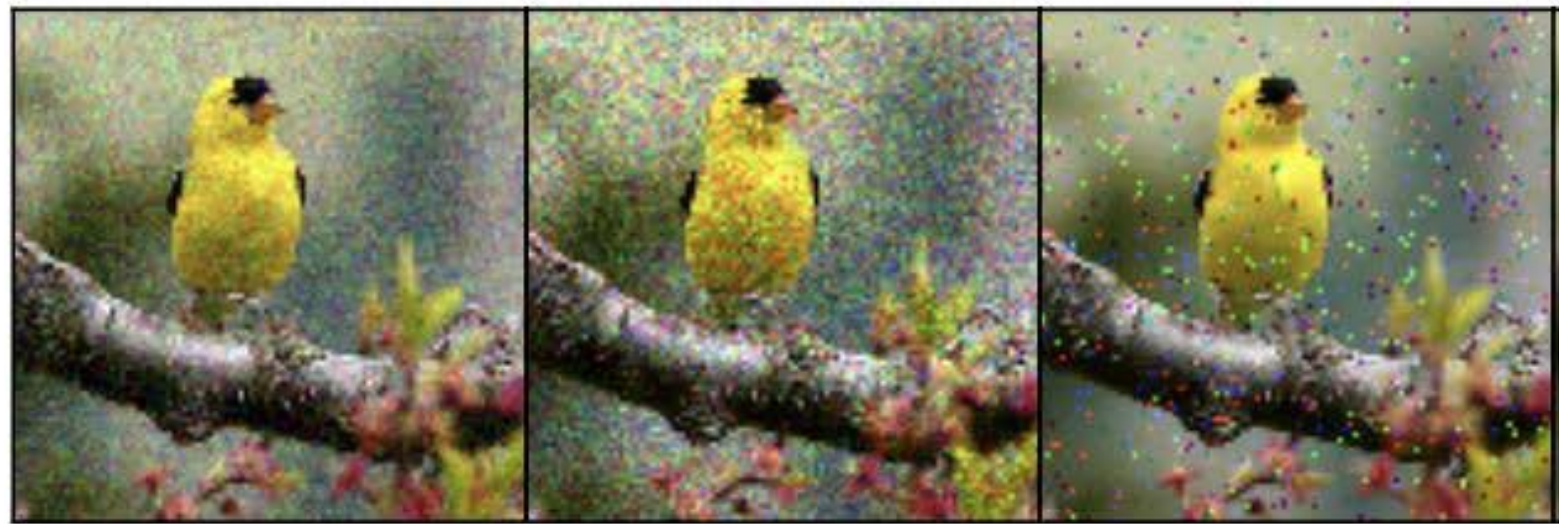

conv 0 (Pruned)
conv 1 (Pruned)
conv 2 (Pruned)
dense 0 (Pruned)
dense 1 (Pruned)
output dense (Pruned)
hls 4 ml
Baseline Pruned
Number of Weights
Weights
0
1000
2000
3000
4000
5000
6000
7000
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

Hooker et al. (2021)

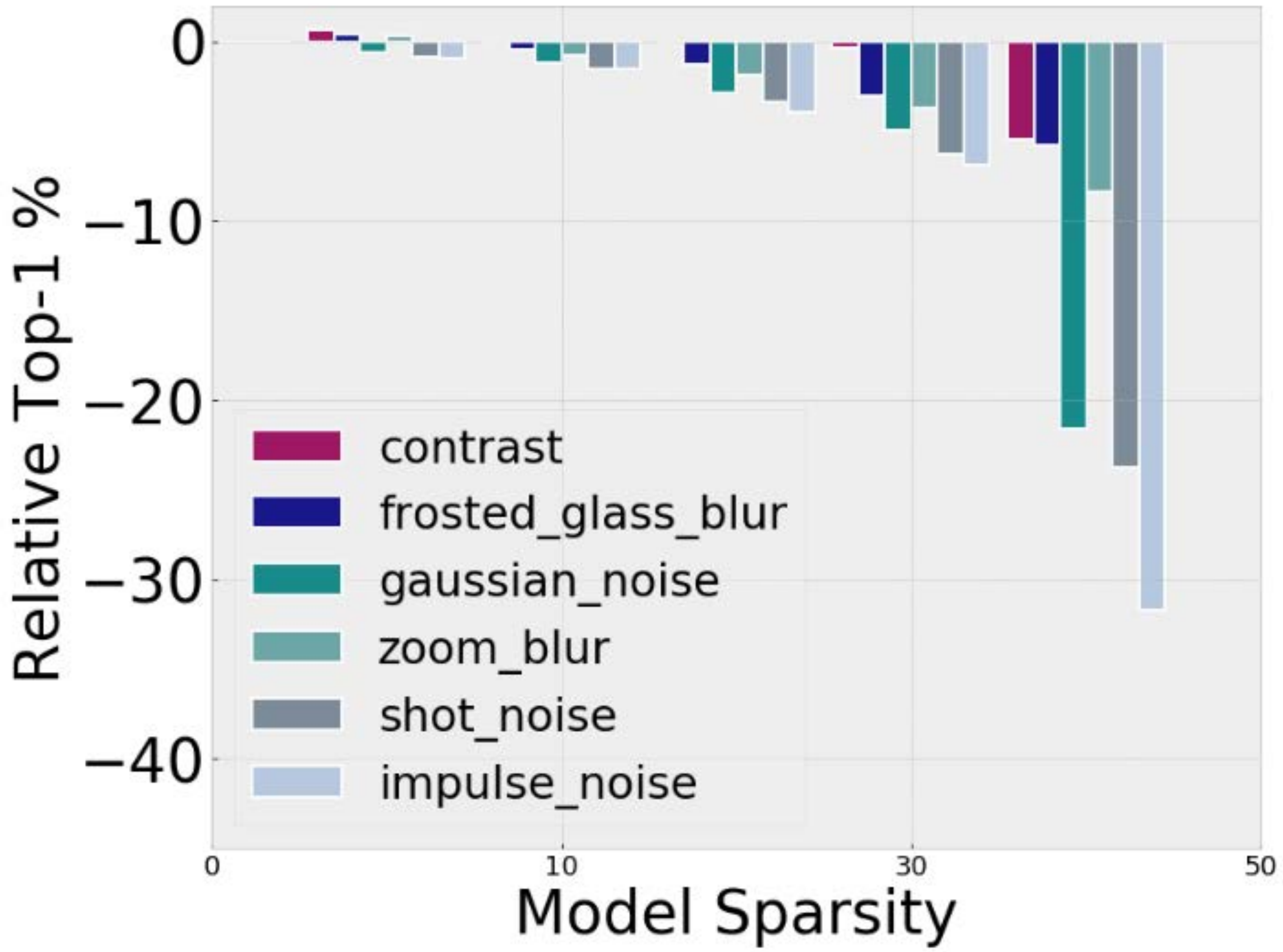
[From Brian Bartoldson]

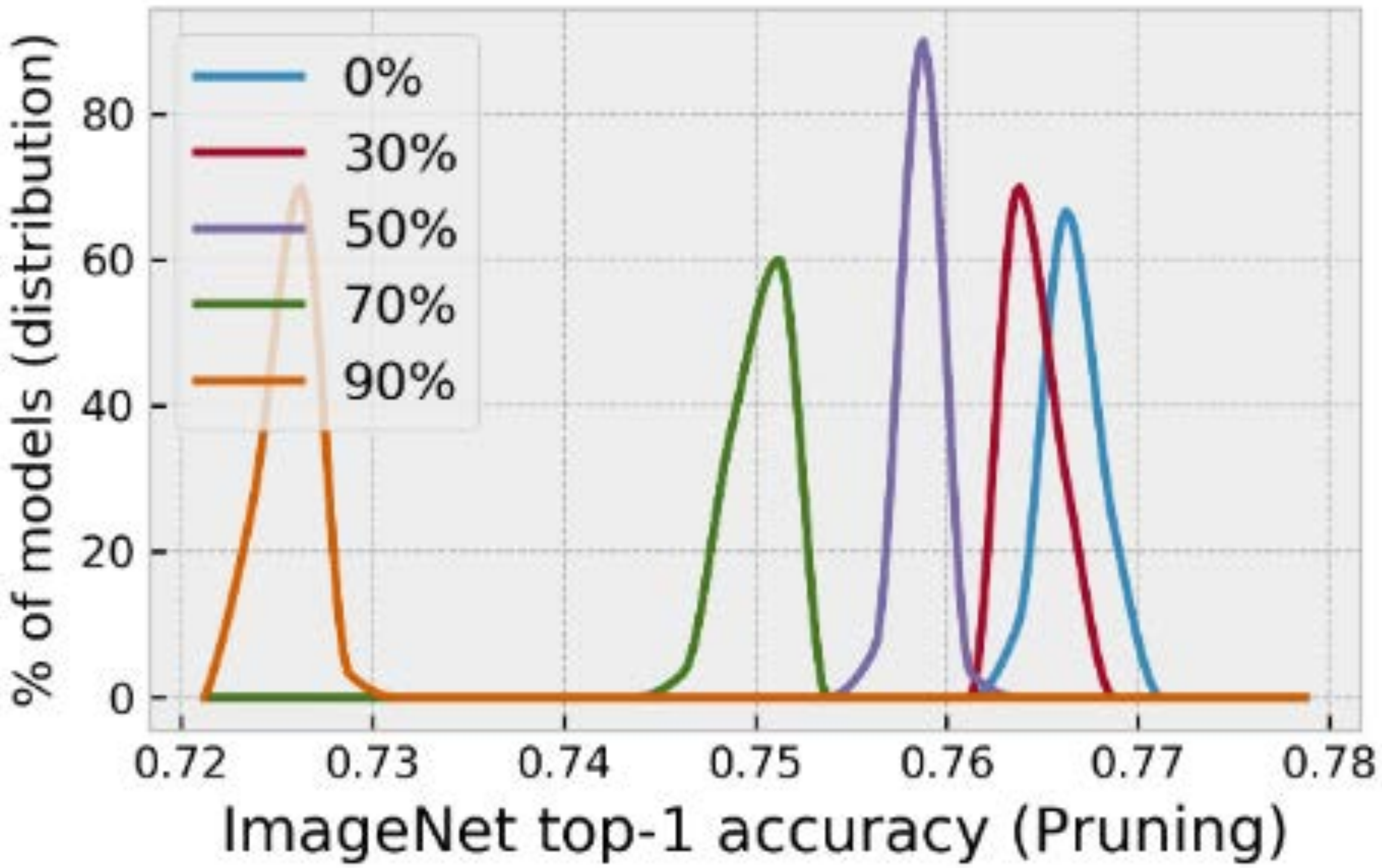
Original image

Gaussian Noise Shot Noise Impulse Noise

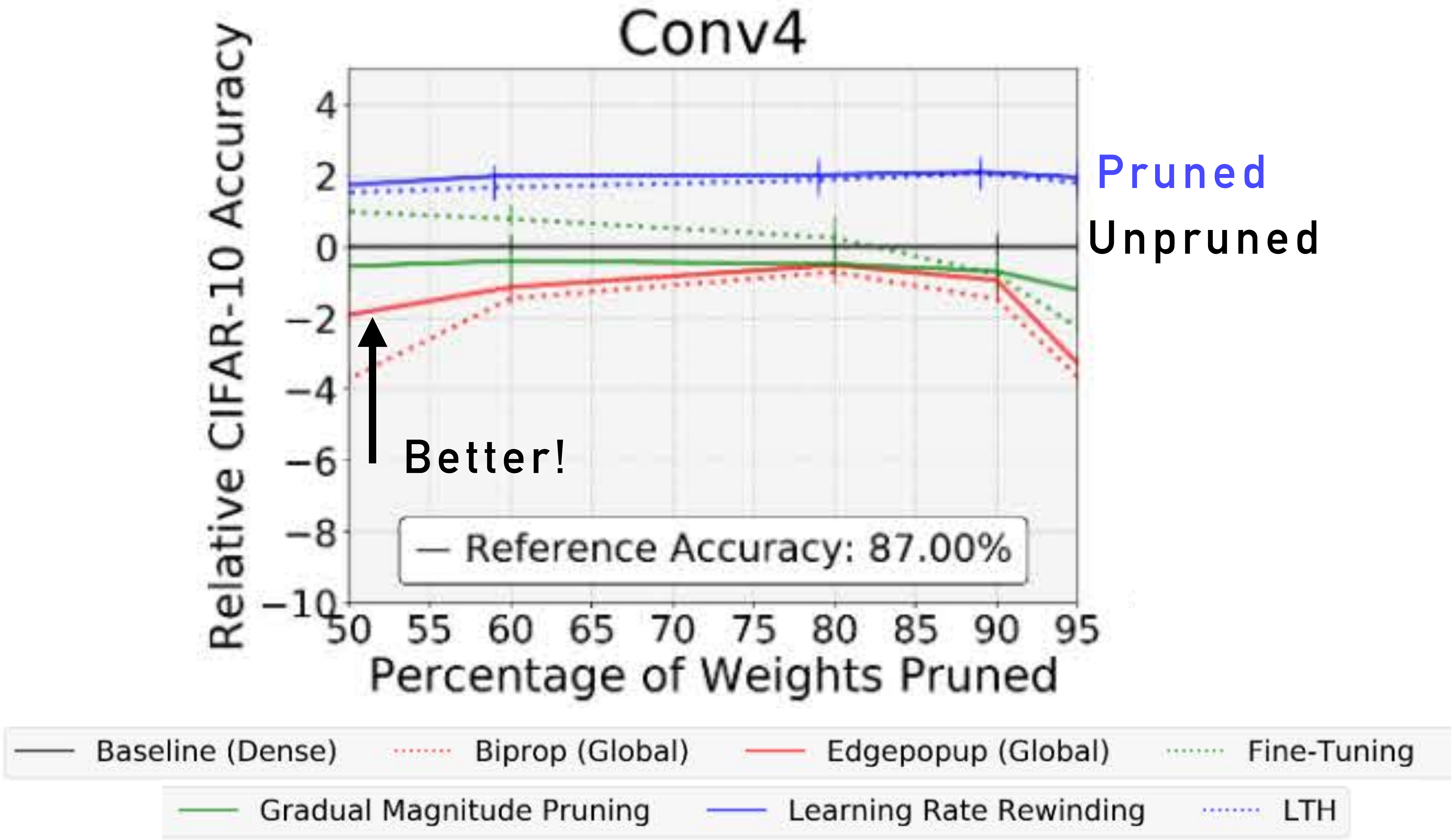


ImageNet-C





From Brian Bartoldson



There exists a optimal network WITHIN each network (lottery ticket)
Uncover it through pruning!

Diffenderfer, Bartoldson, et al. (2021)

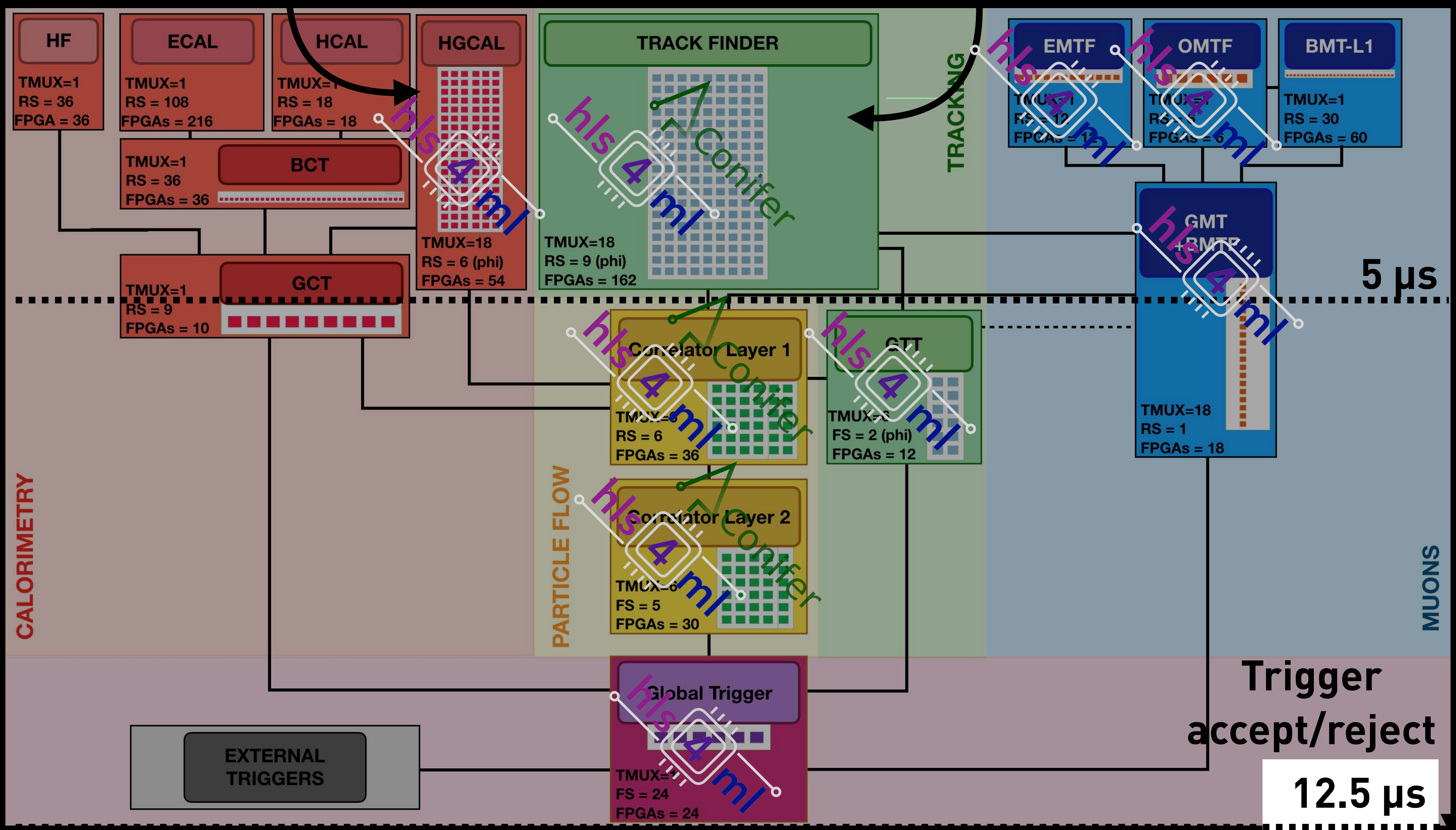
hls4ml

[hls4ml tutorial]

Nanosecond ML inference on FPGAs!

40 billion inferences/s during HL-LHC

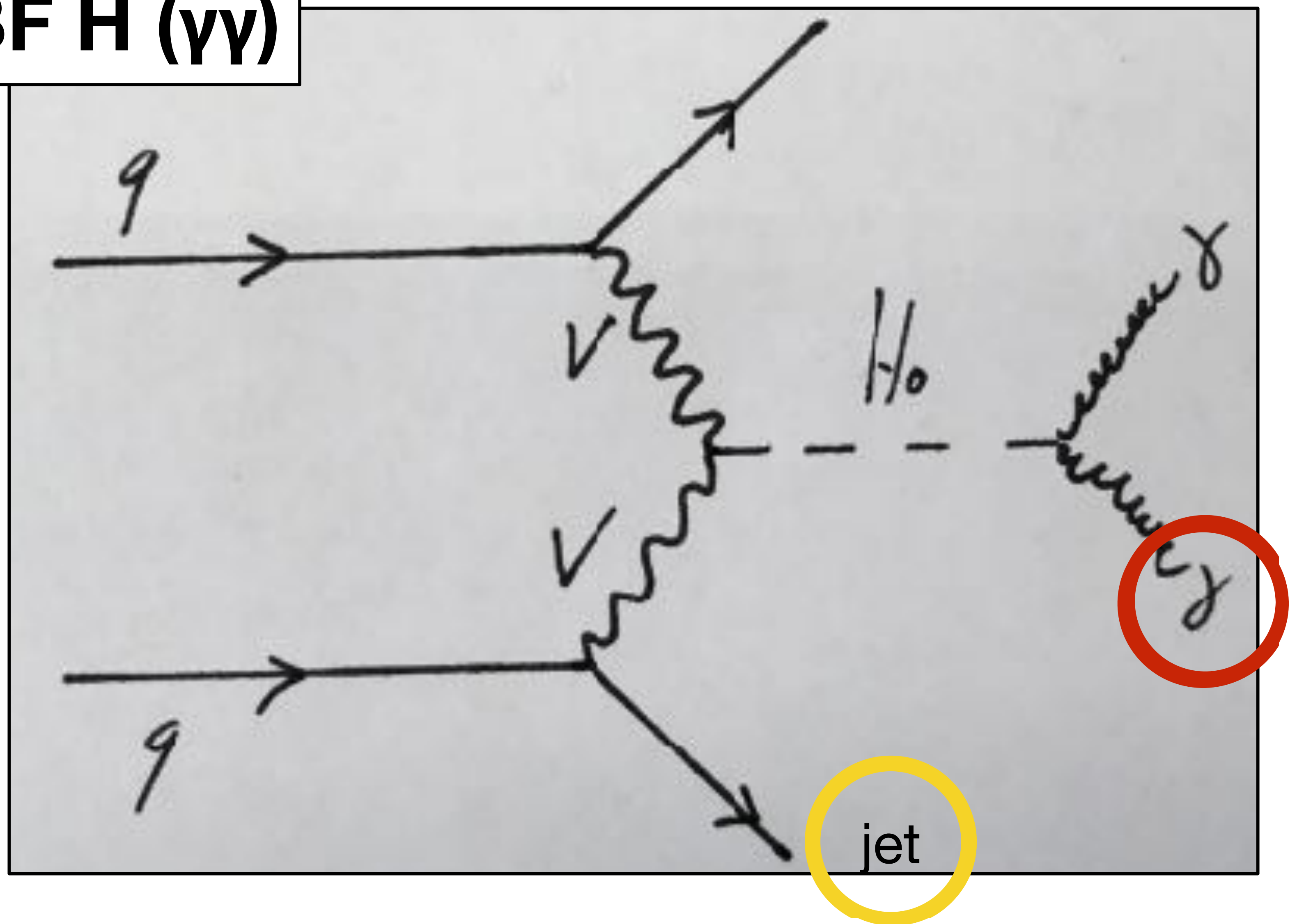
(≈ all inferences at Google)



HEP developed libraries for fast ML on FPGAs

VBF H (γγ)
jet
200 vertices
+
Layer 5
CE-E
Layer 5
y(mm)
x(mm)
E(MIP)
Layer 30
CE-H
Layer 30
y(mm)
x(mm)
E(MIP)

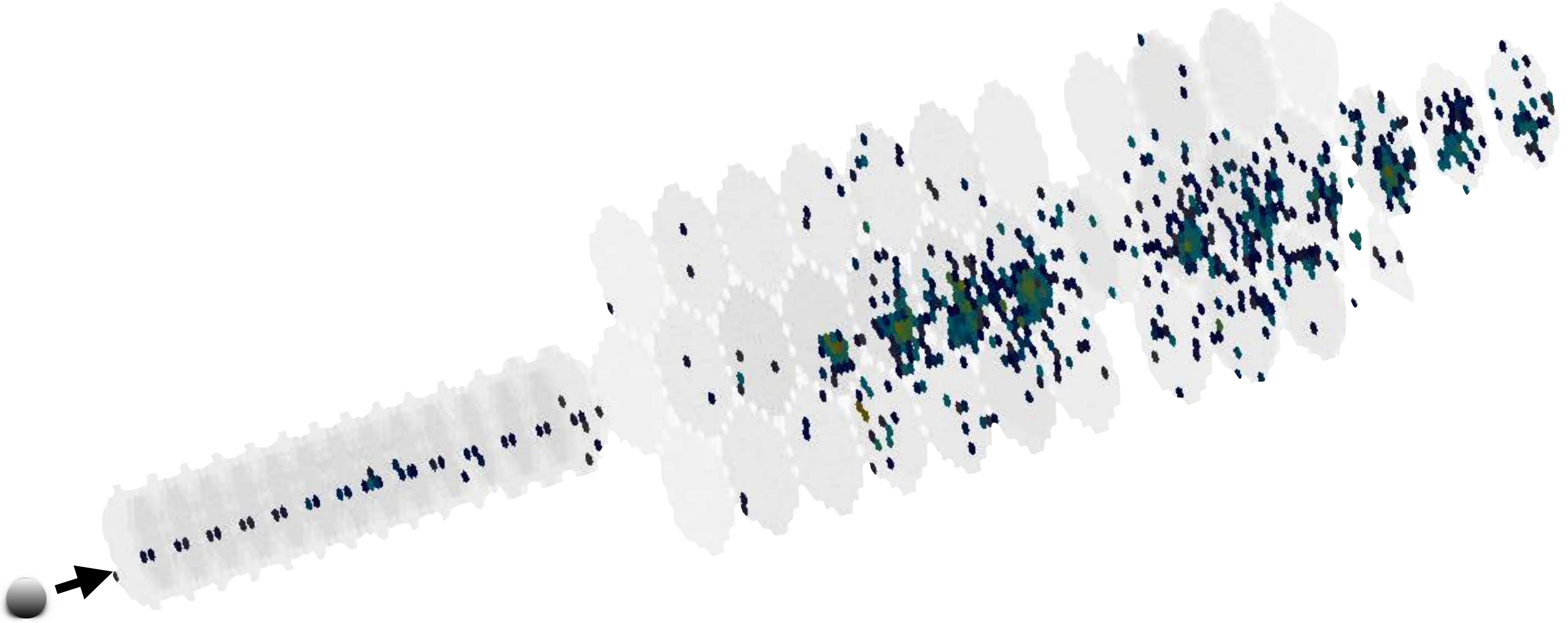
VBF H (γγ)

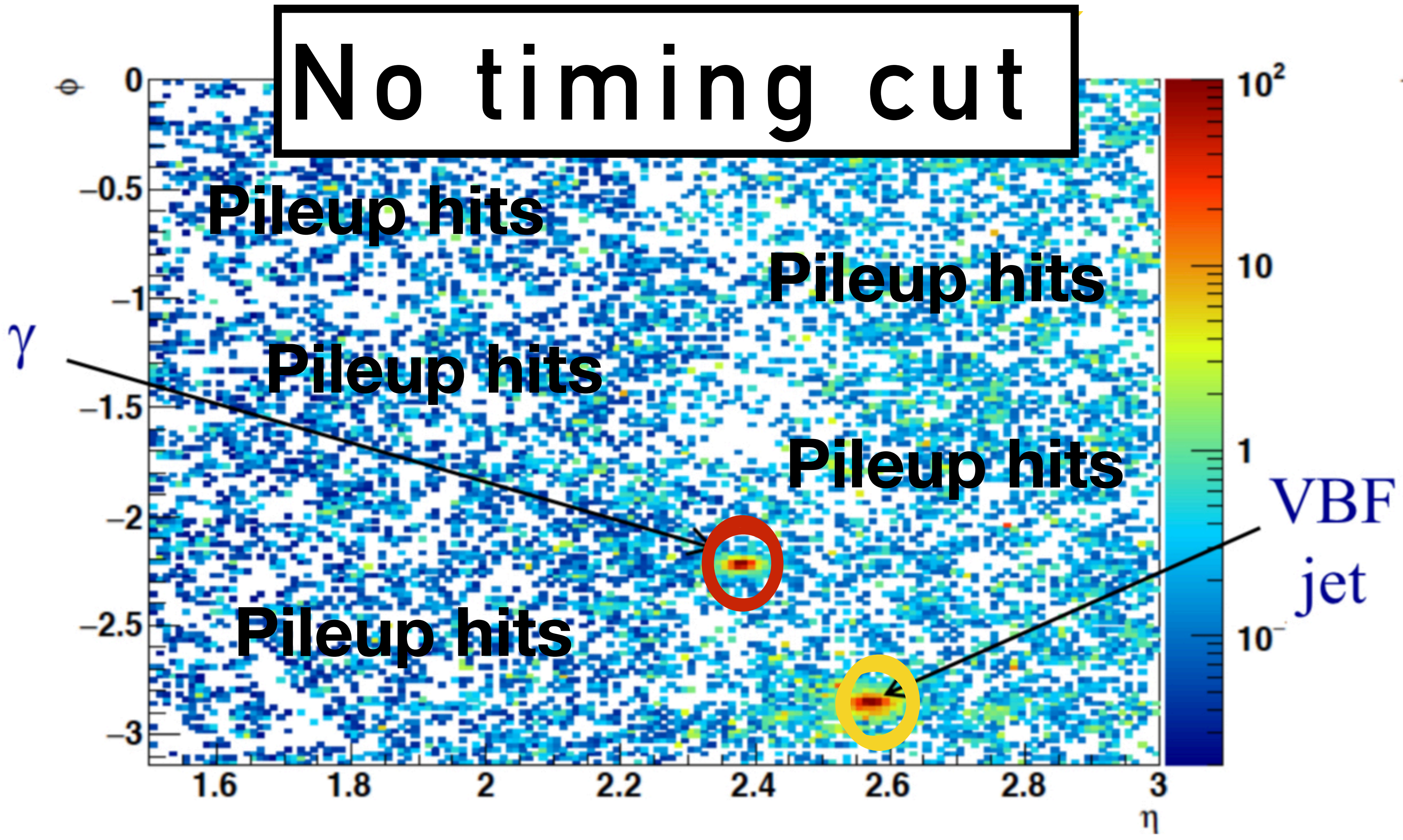


q
q
V
V
Ho
γ
γ
jet

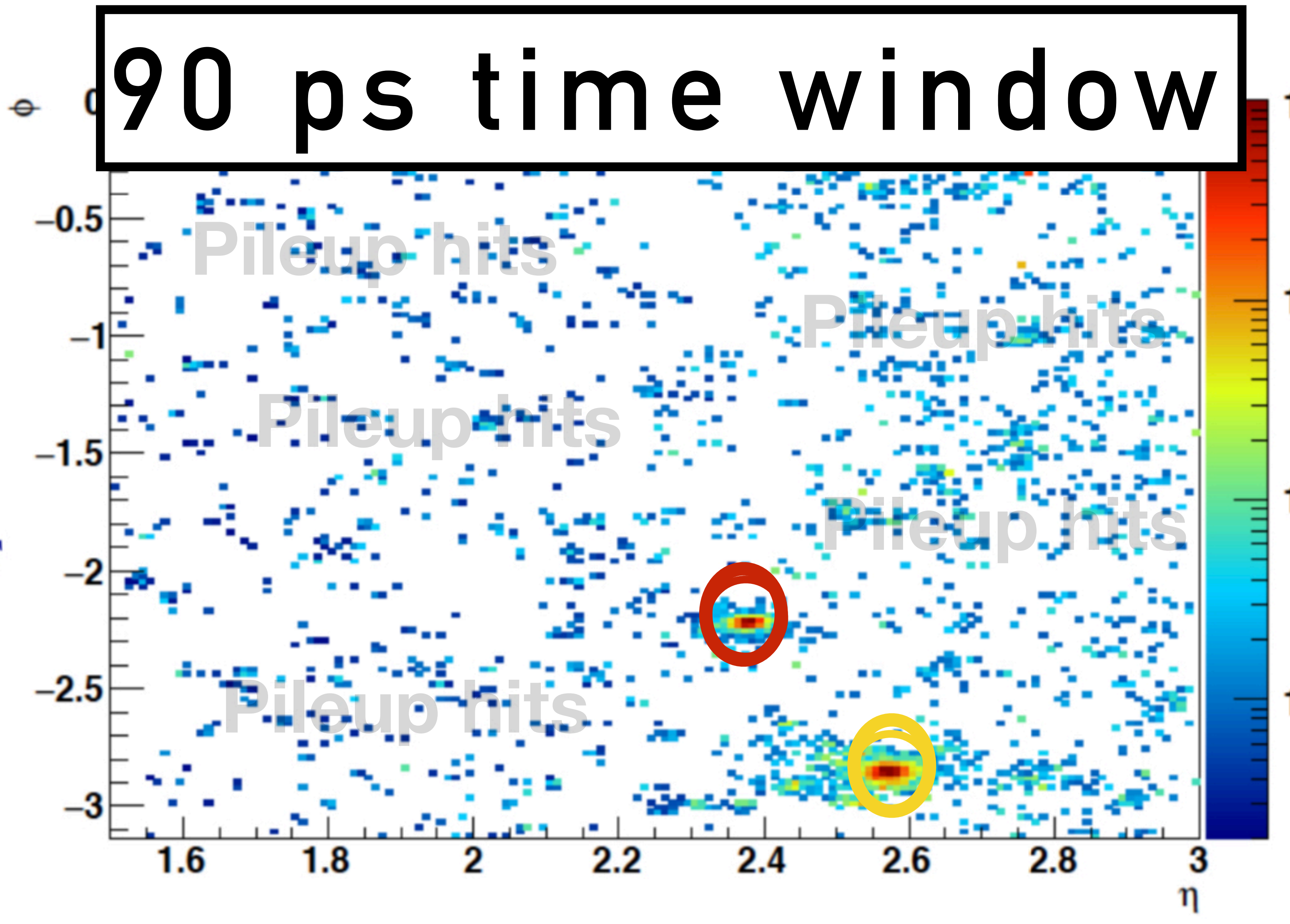
+

200 vertices



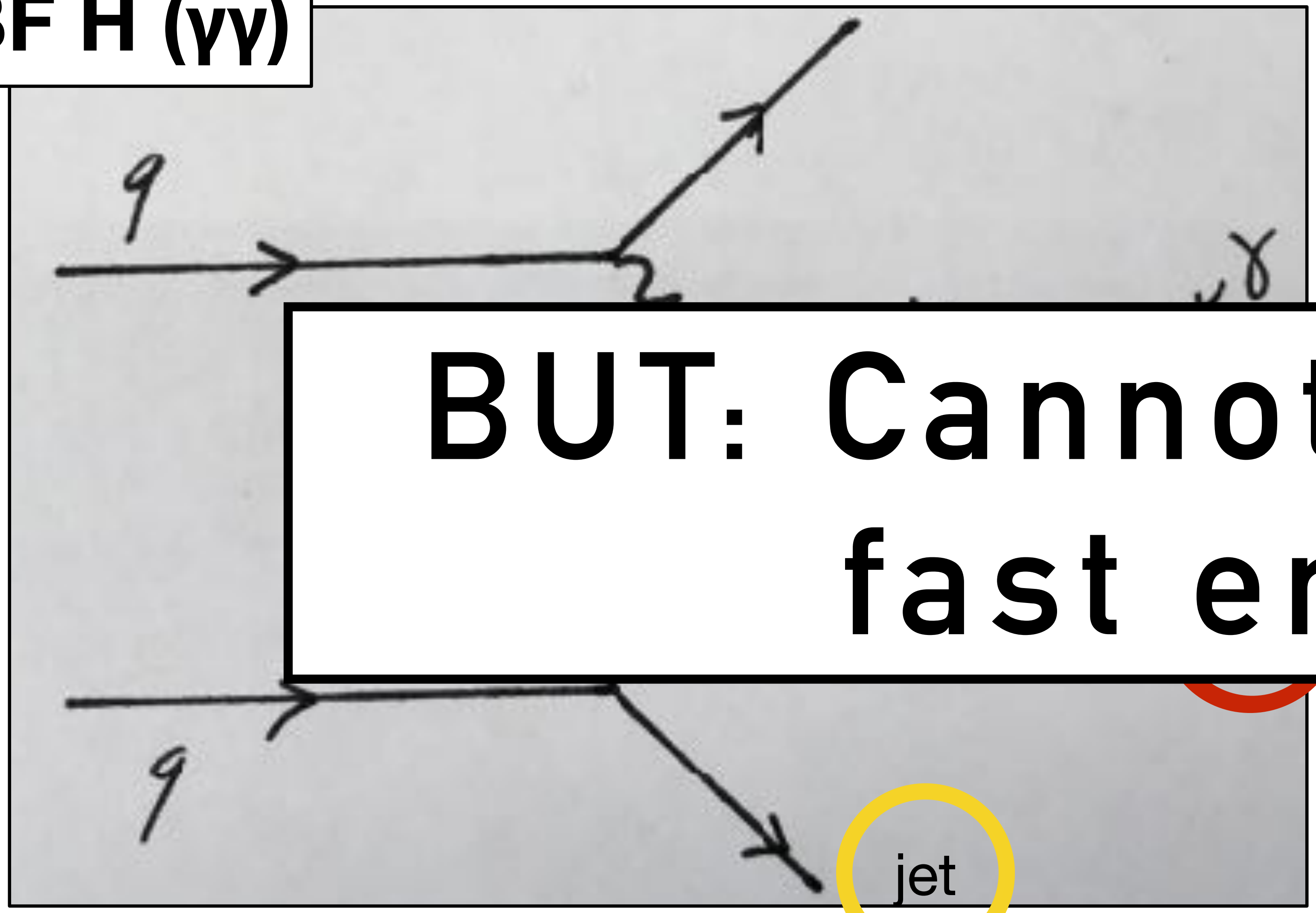


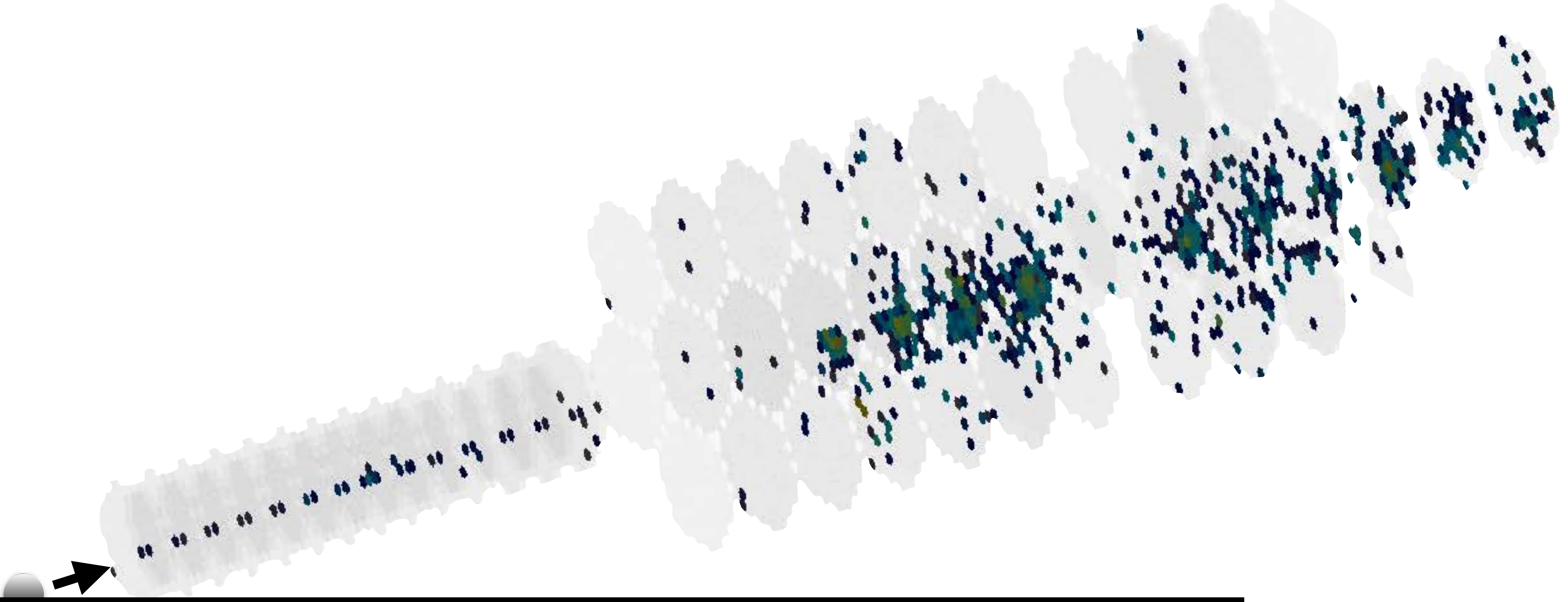
No timing cut
Pileup hits
Pileup hits
Pileup hits
Pileup hits
Pileup hits
γ
VBF jet
φ
η



90 ps time window
Pileup hits
Pileup hits
Pileup hits
Pileup hits
Pileup hits
φ
η

VBF H (γγ)

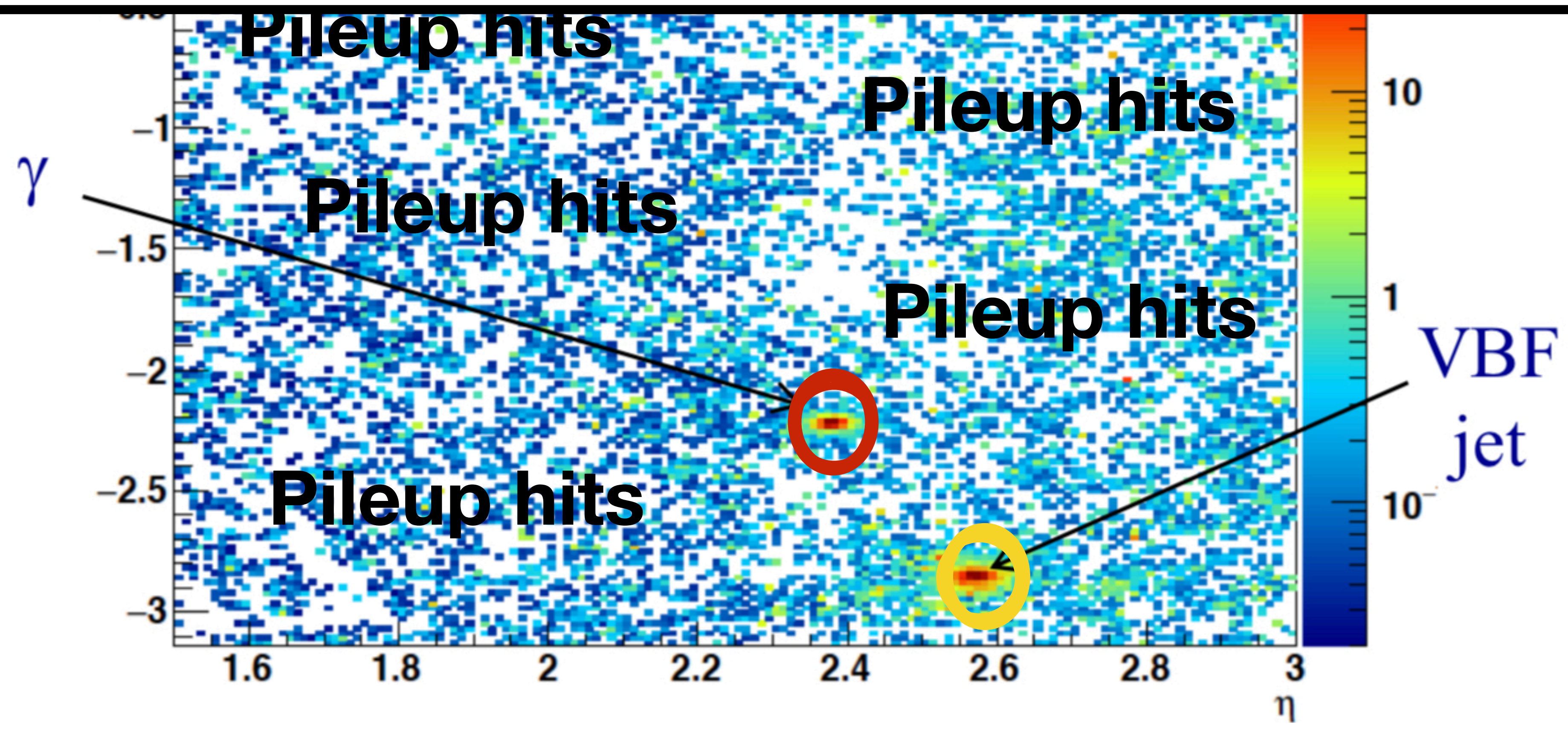


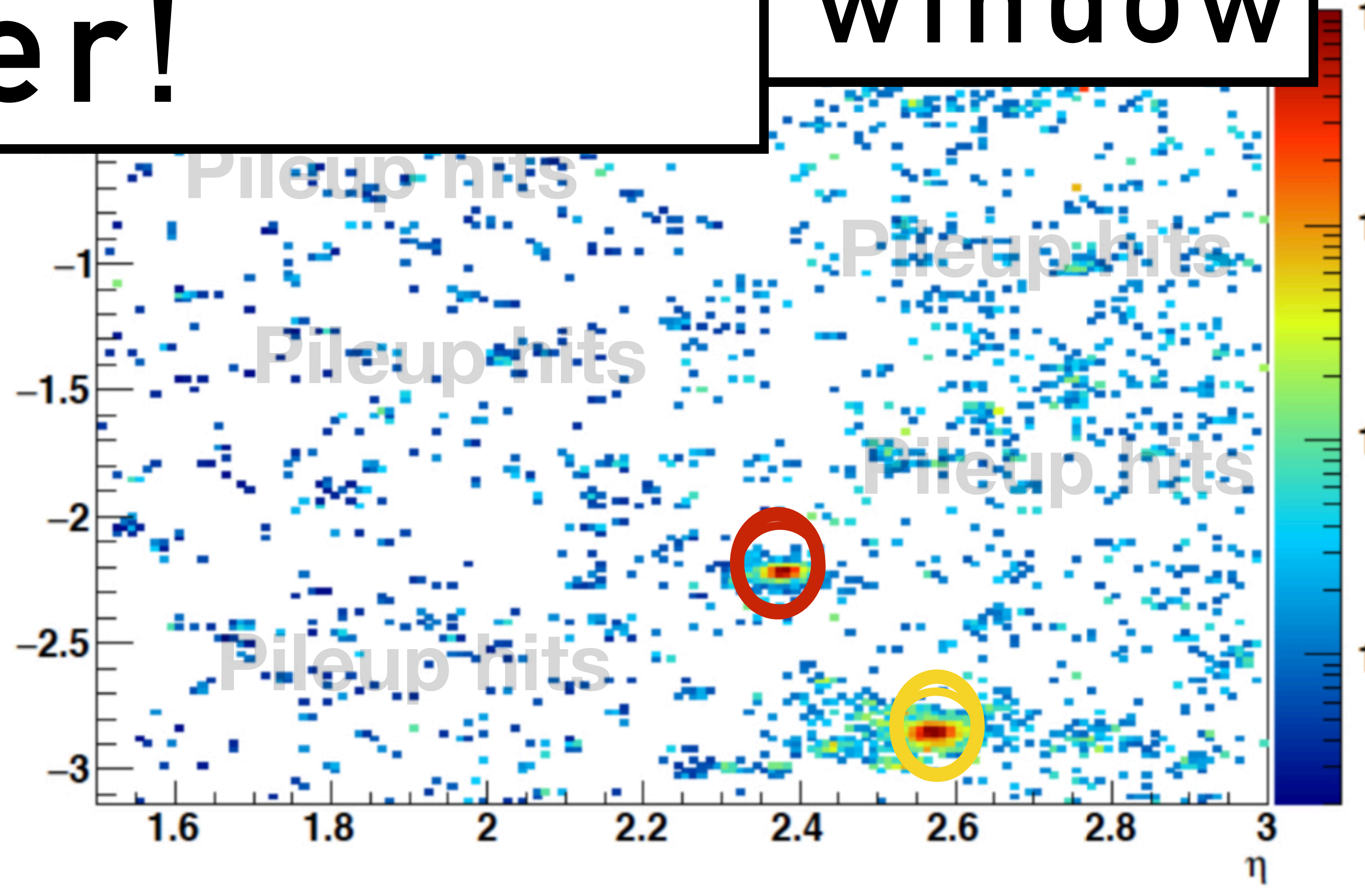


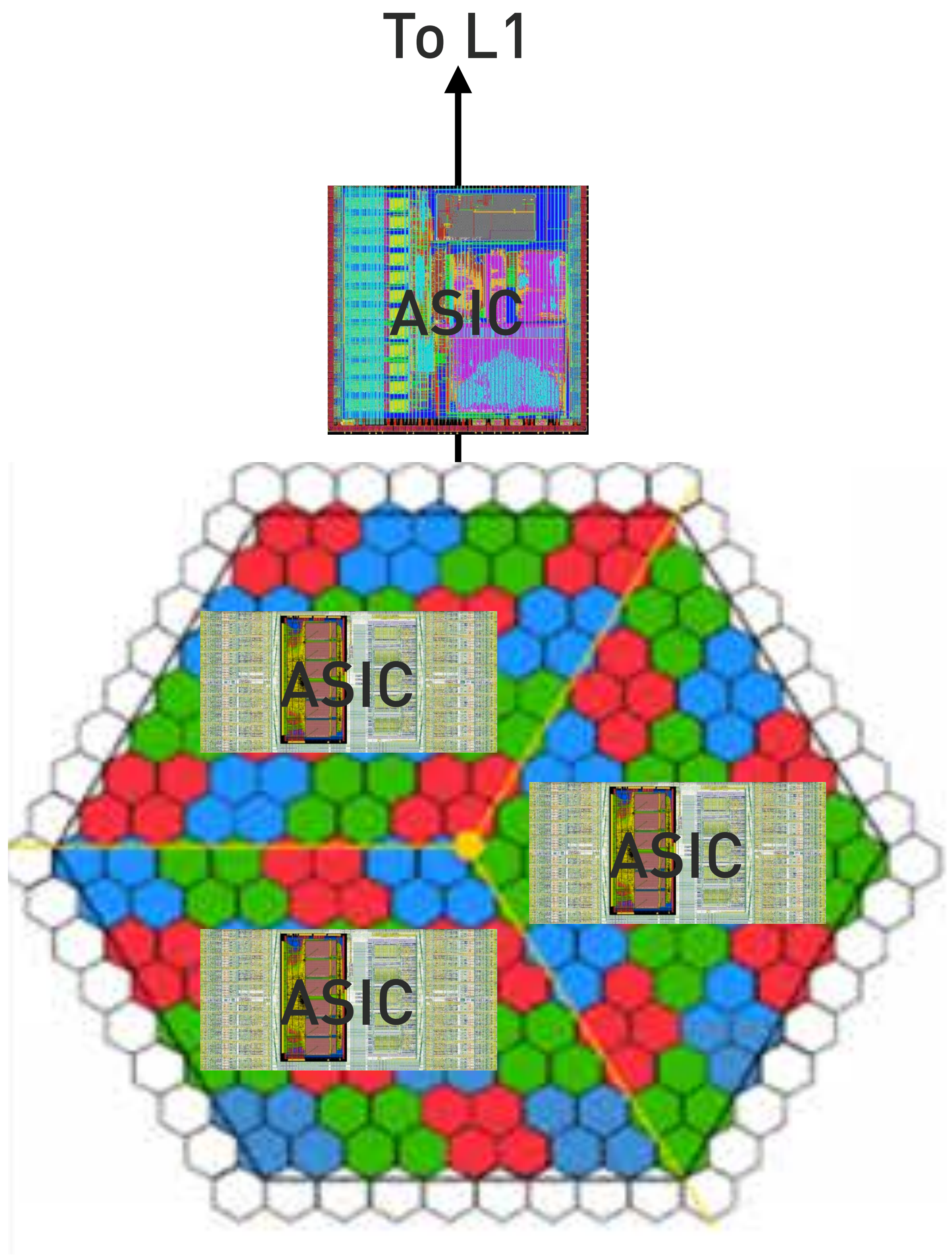
BUT: Cannot read out all these channels fast enough for L1 to trigger!

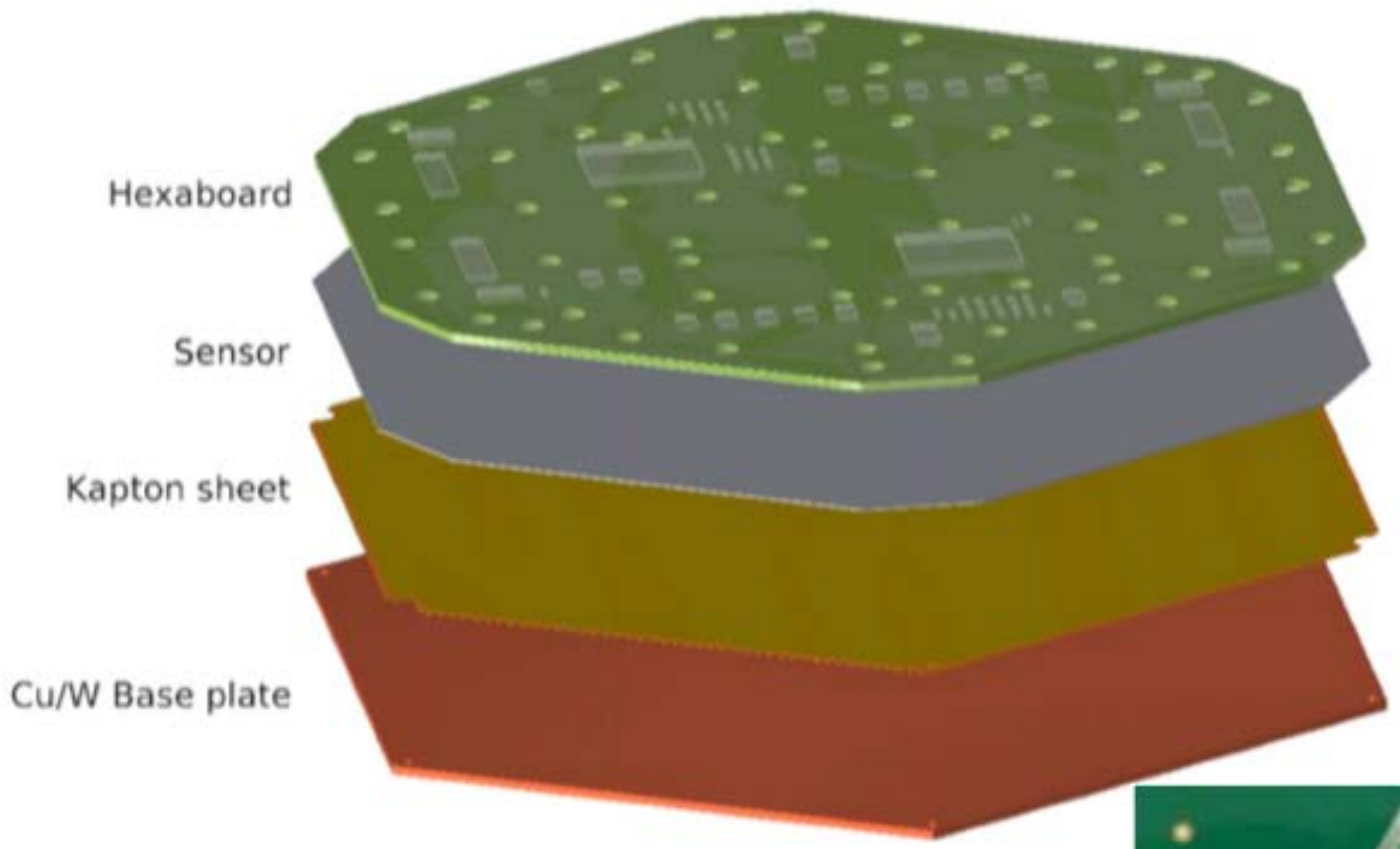
window

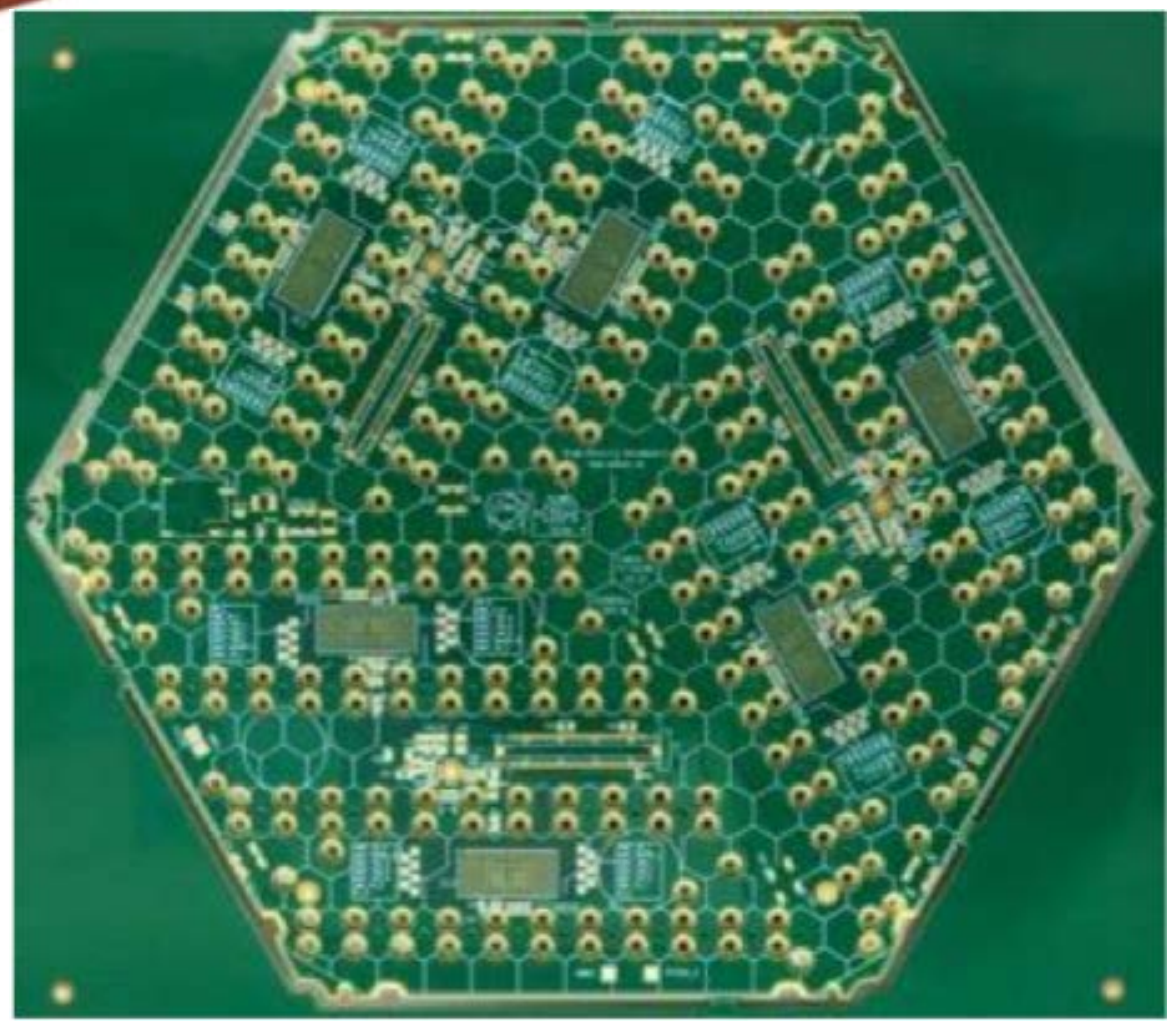
+

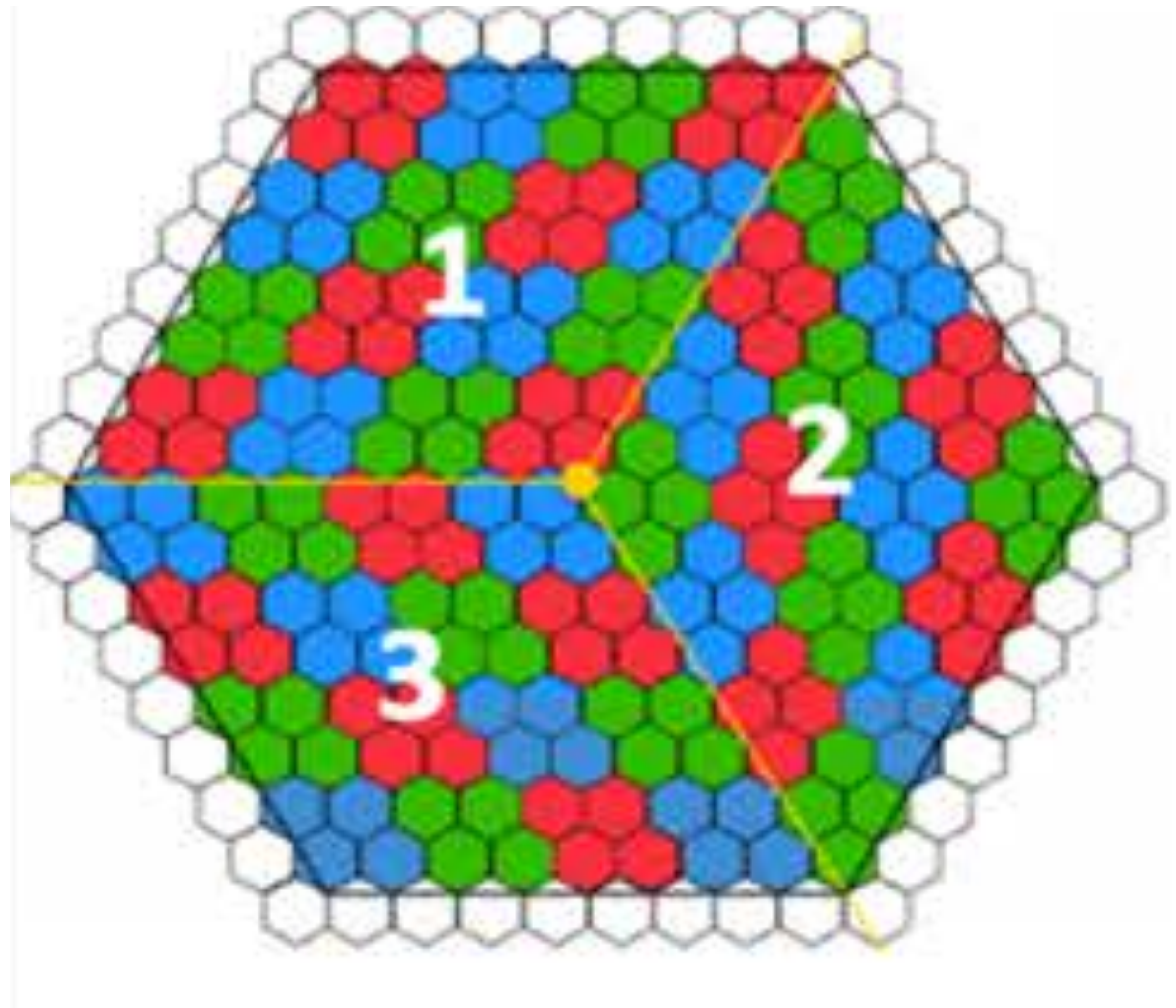


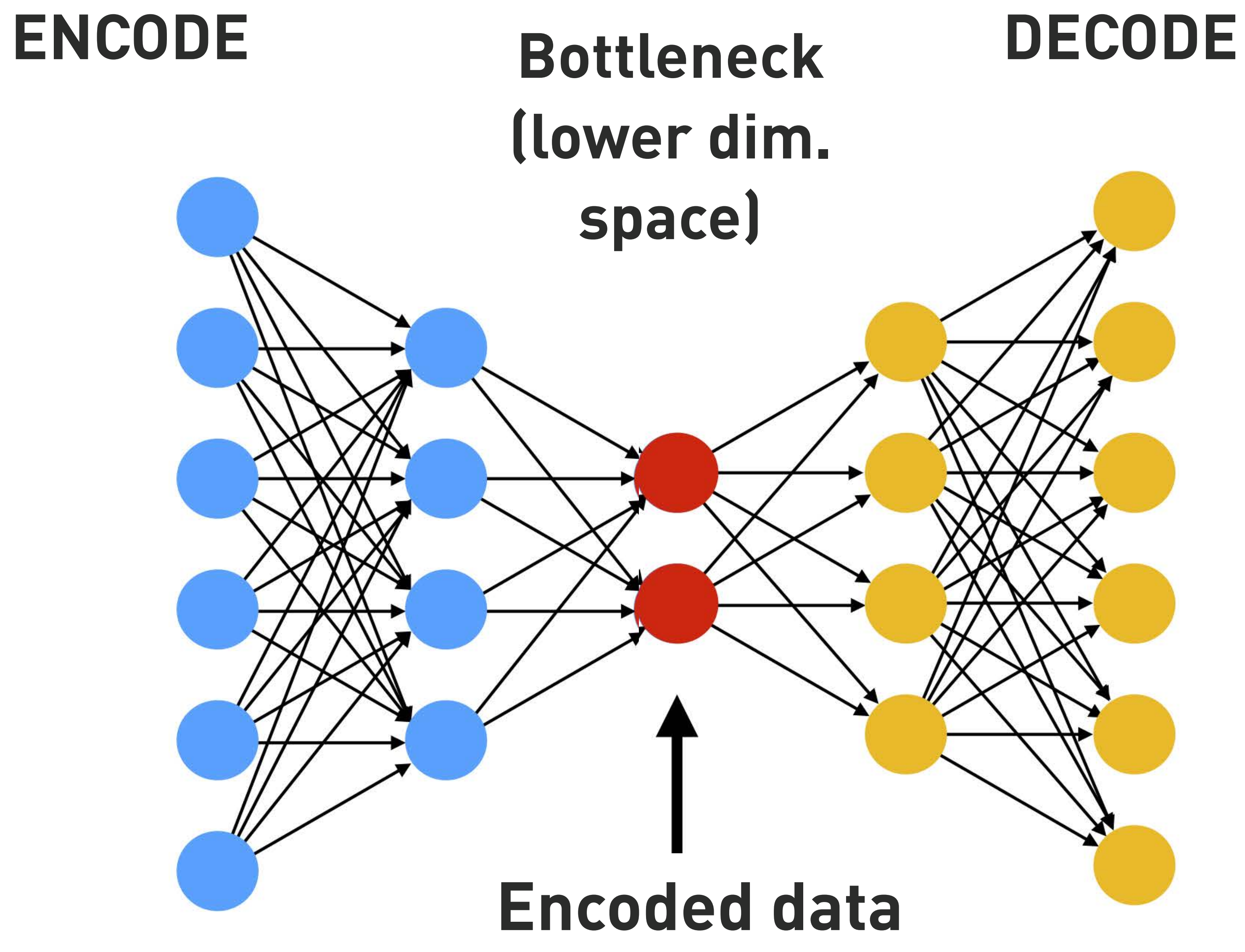


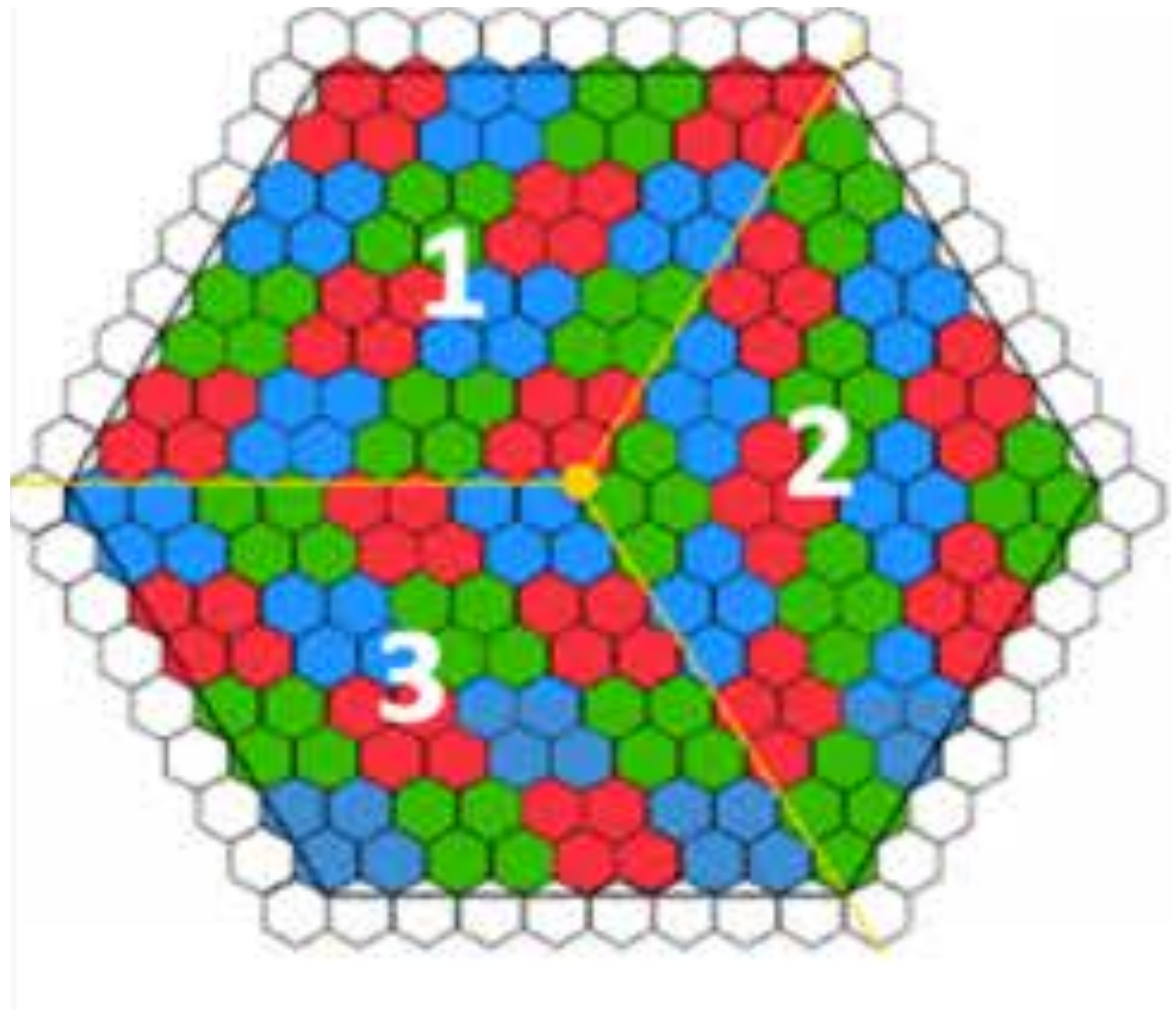

To L1
ASIC
ASIC
ASIC
ASIC


Hexaboard
Sensor
Kapton sheet
Cu/W Base plate





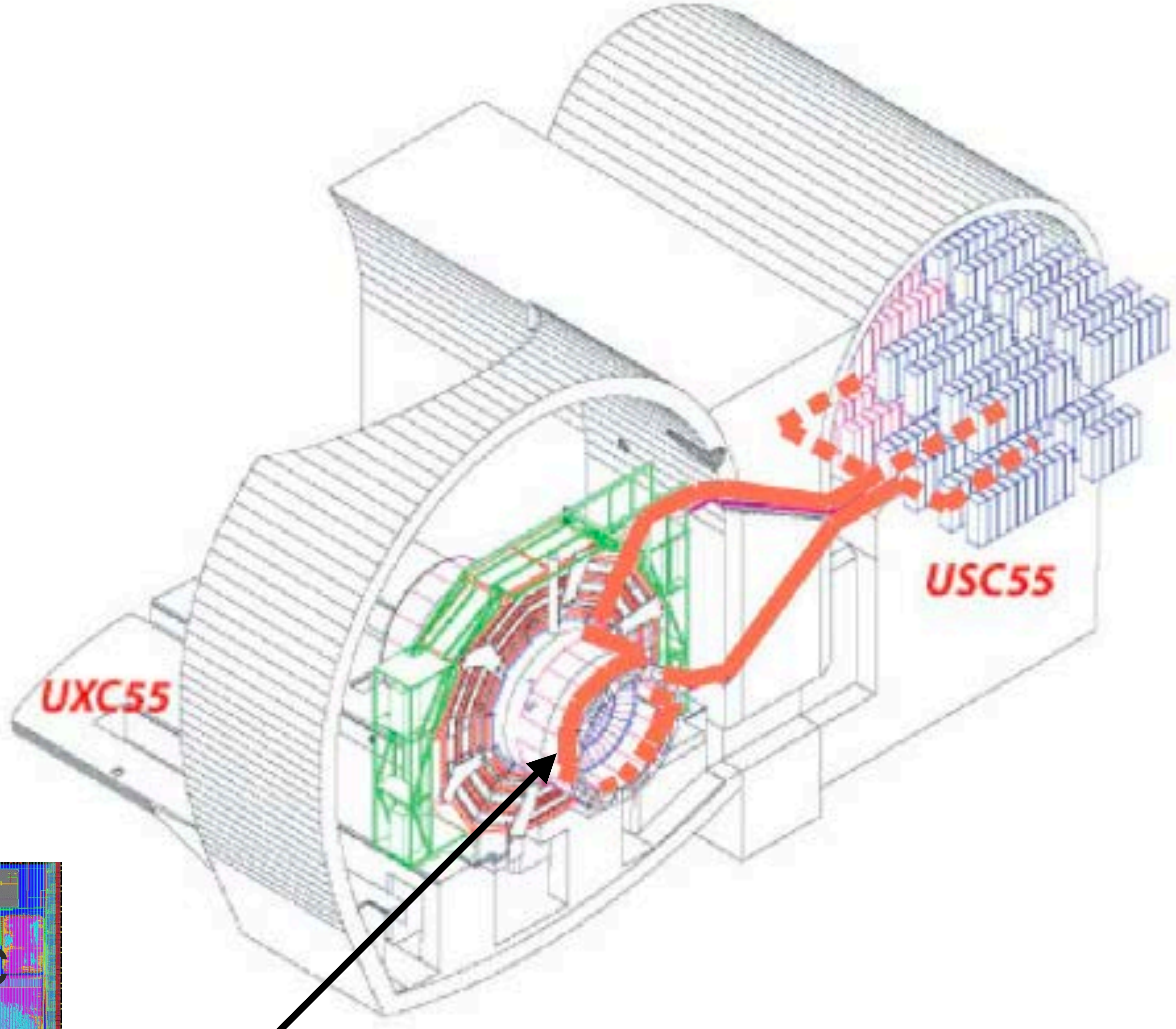


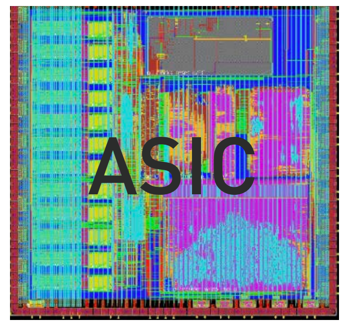


Variational Autoencoder

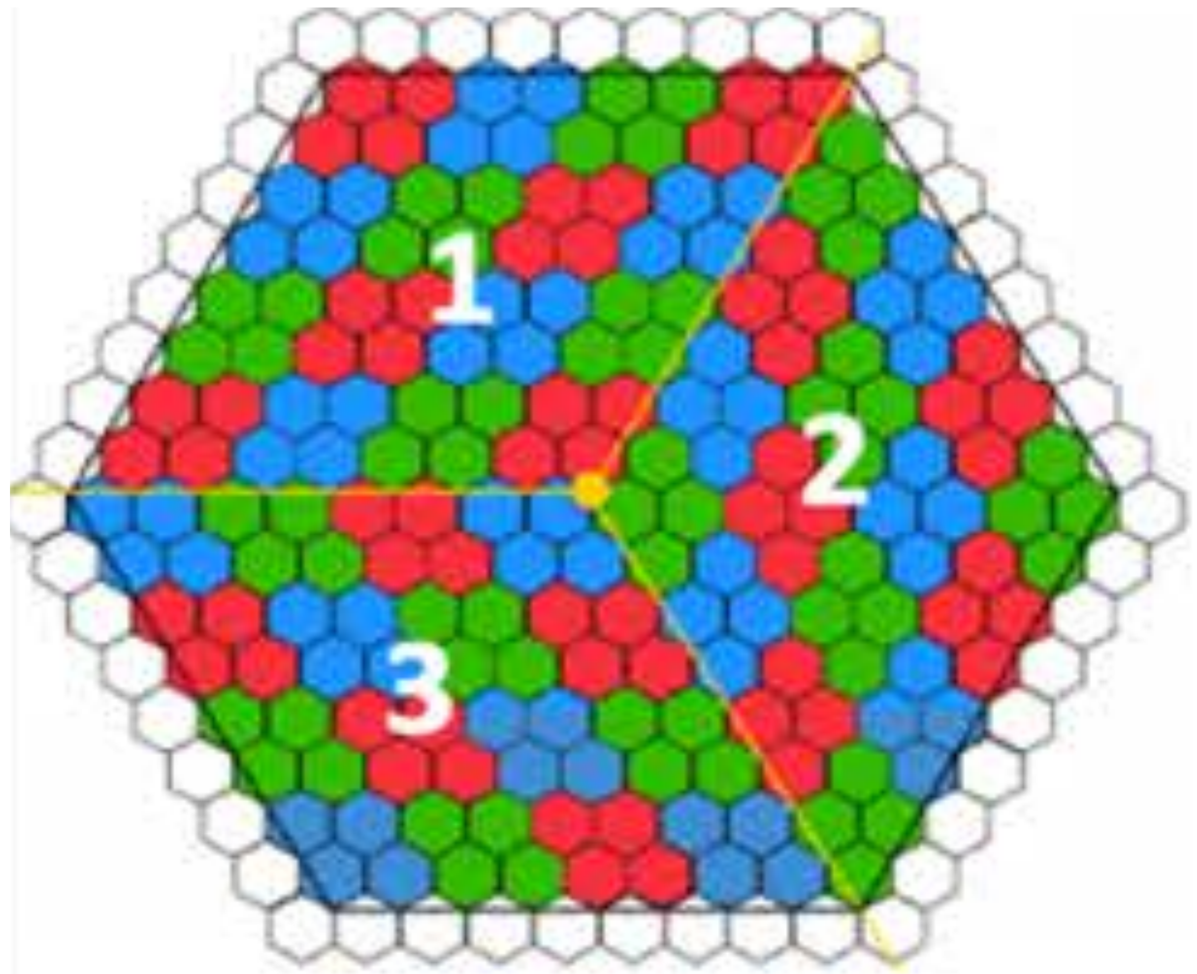
__ECON-T, D. Noonan__

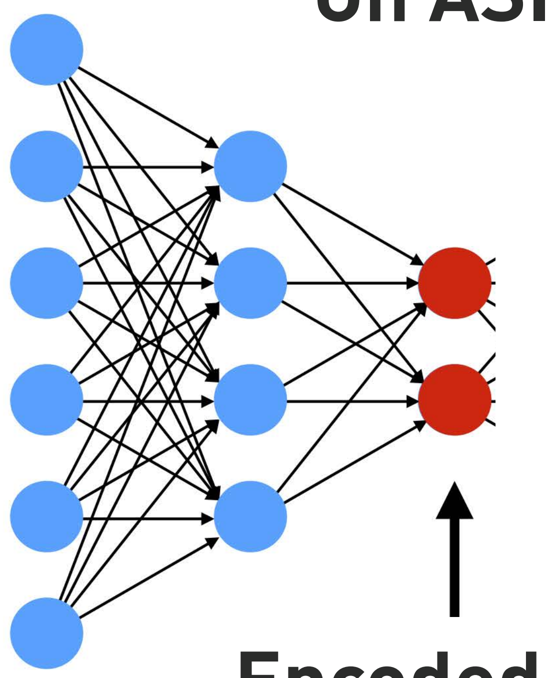
__AEs for compression also at LHCb!__



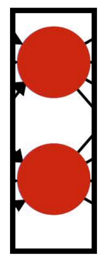


On ASIC





Encoded data



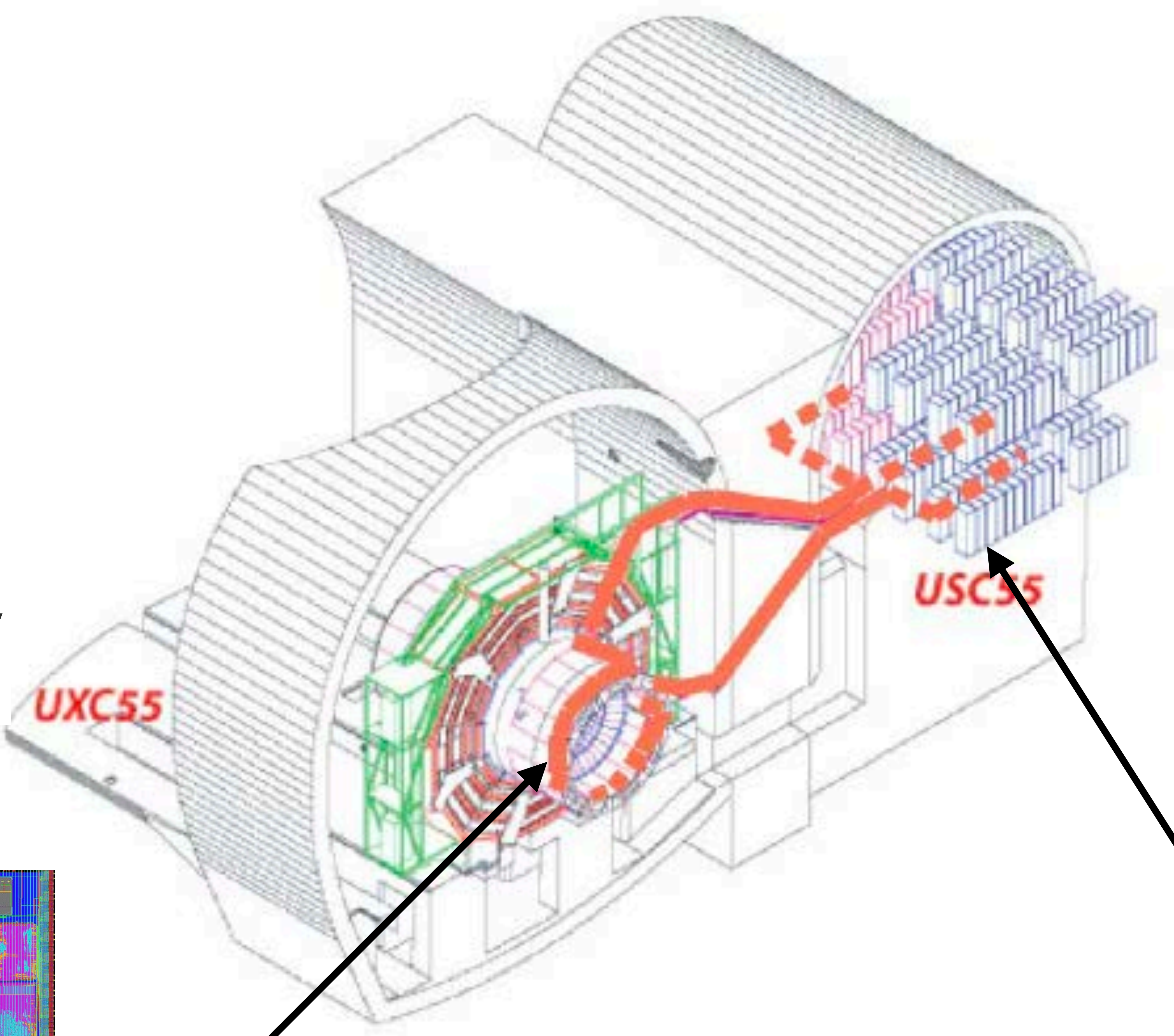
Transmit encoded data!

ECON-T, D. Noonan

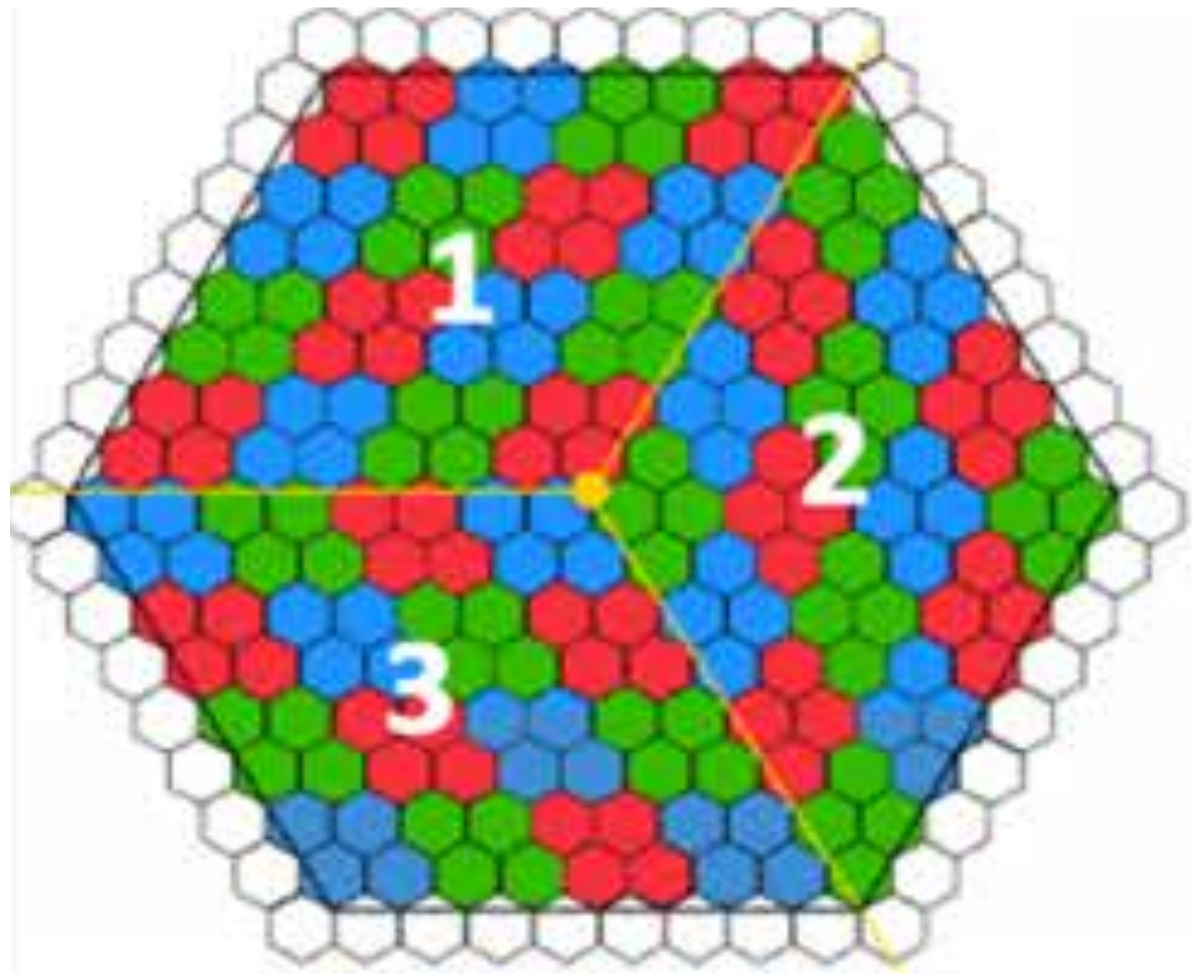
AEs for compression also at LHCb!

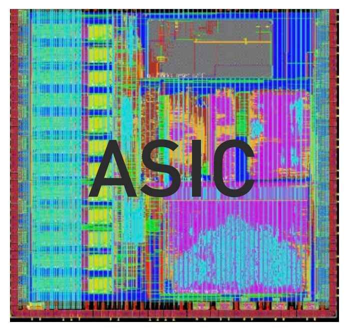
<u>*ECON-T, D. Noonan*</u>

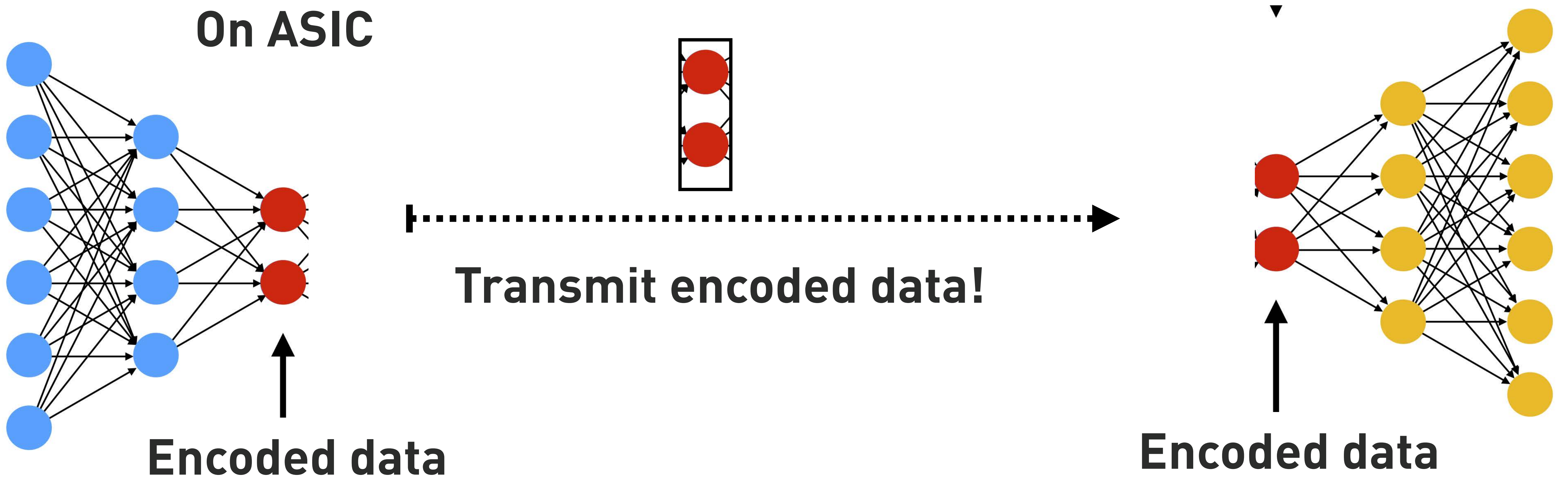
<u>*AEs for compression also at LHCb!*</u>

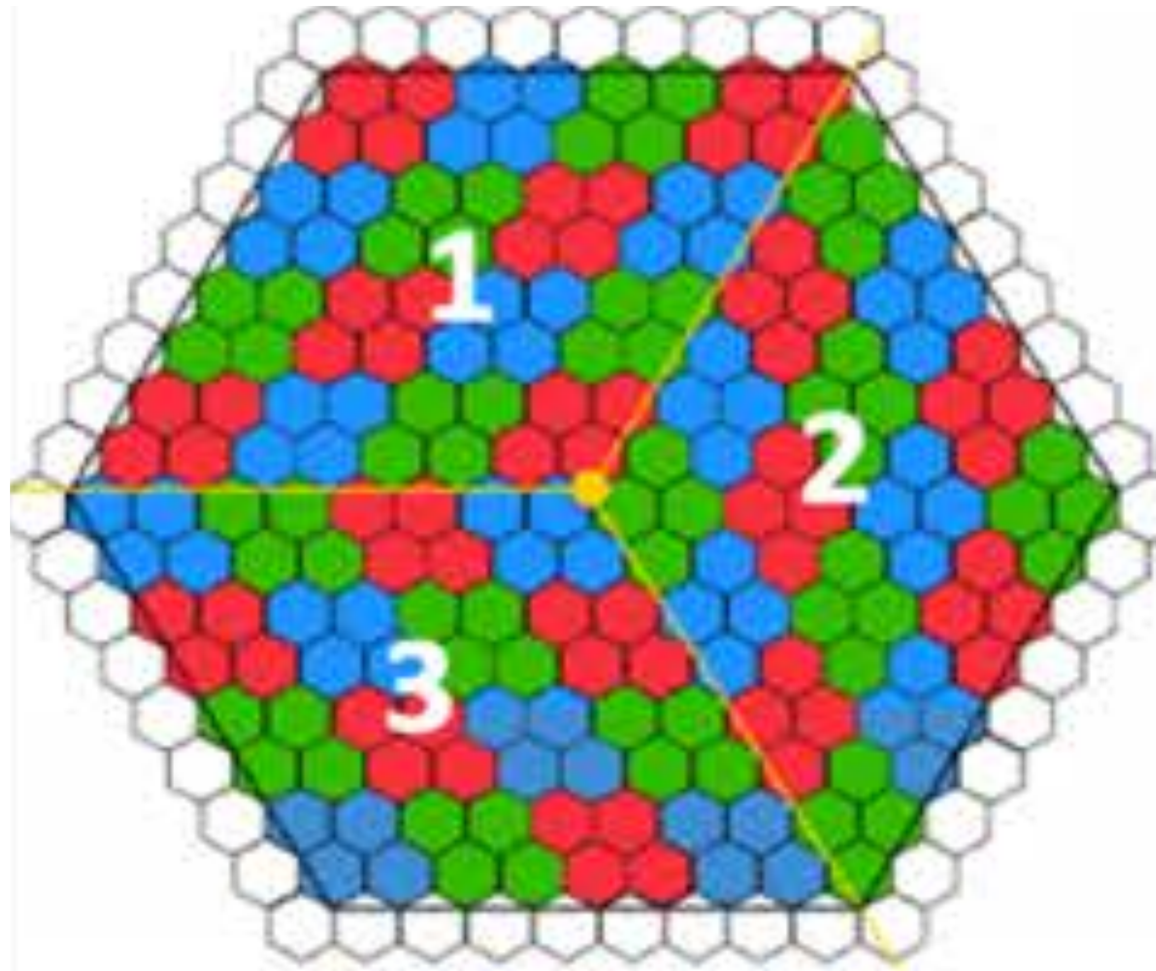


- 75-100 mW
- Triplicated w/b for radiation safety
  Reprogrammable w/b over IC2!



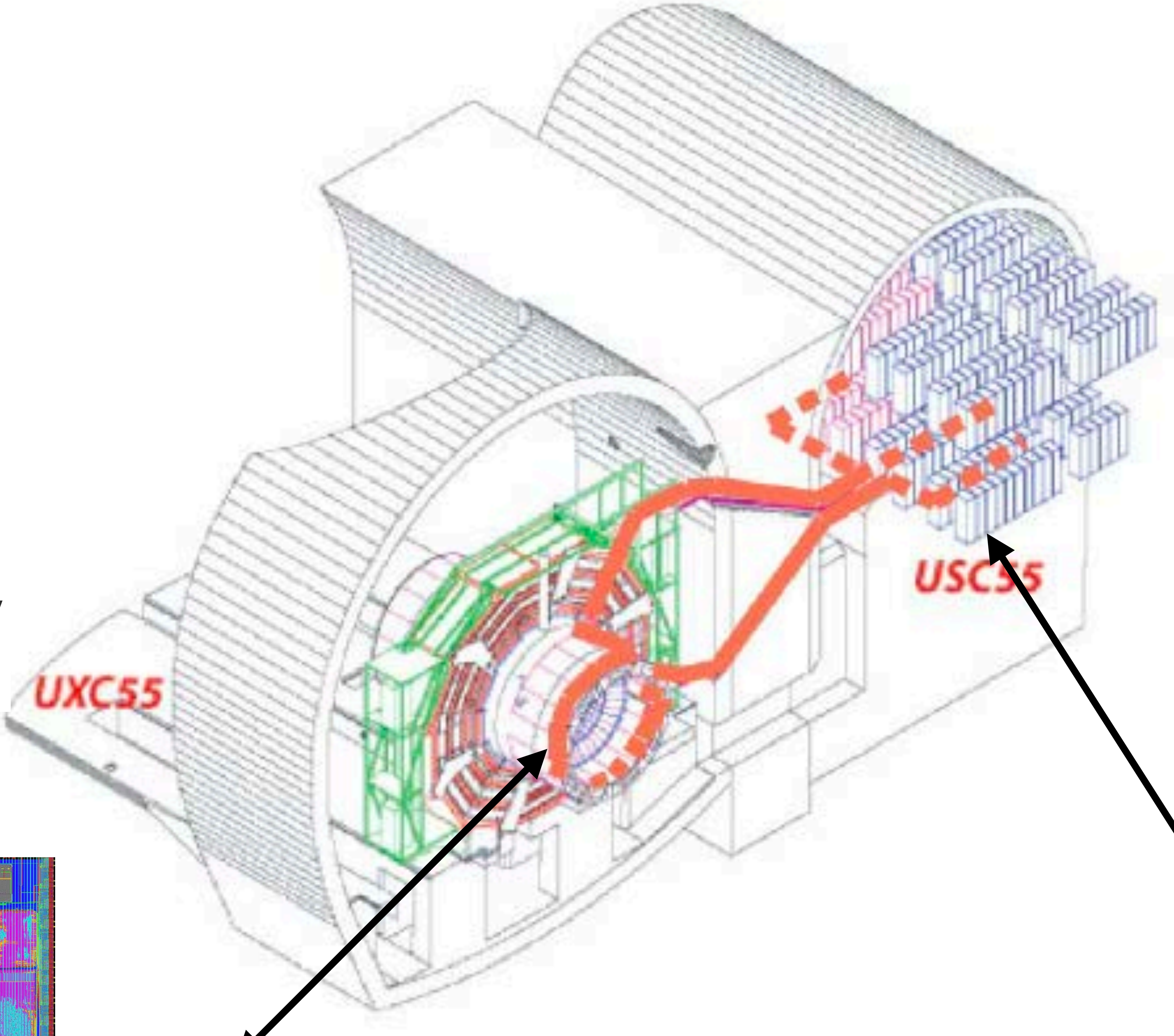




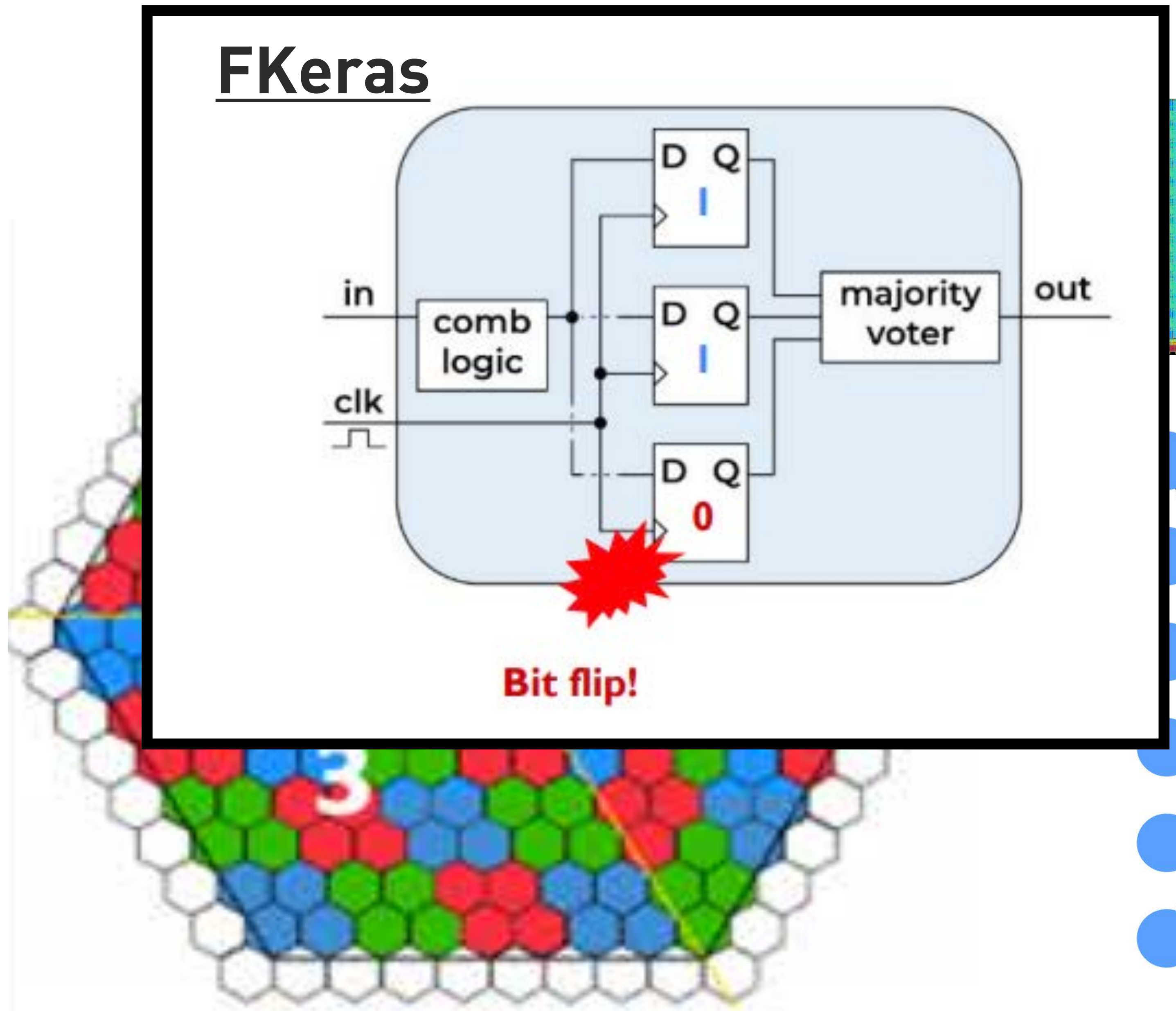


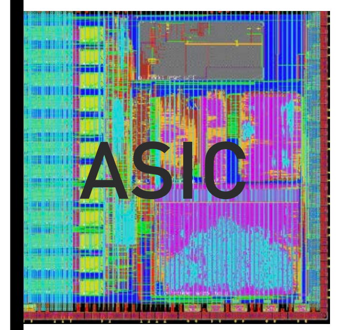
__ECON-T, D. Noonan__

__AEs for compression also at LHCb!__

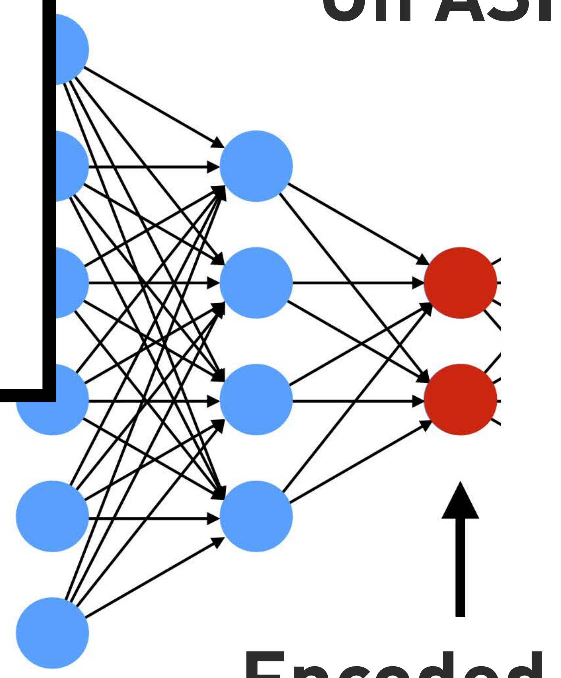


- **75-100 mW**
- **Triplicated w/b for radiation safety**
  Reprogrammable w/b over IC2!

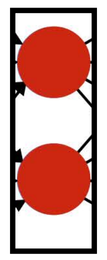




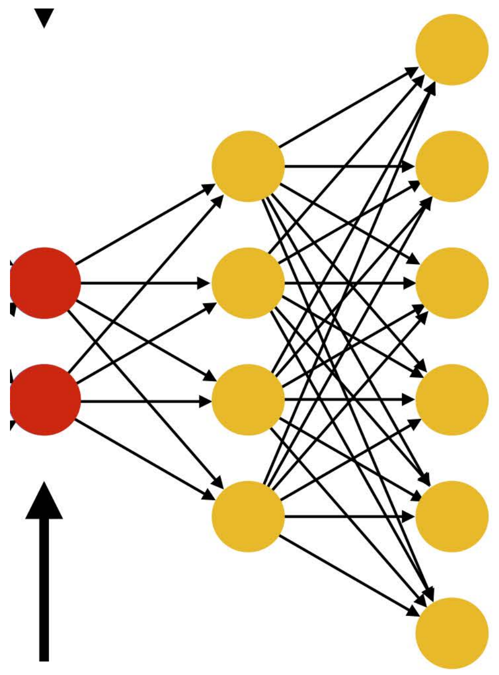
On ASIC



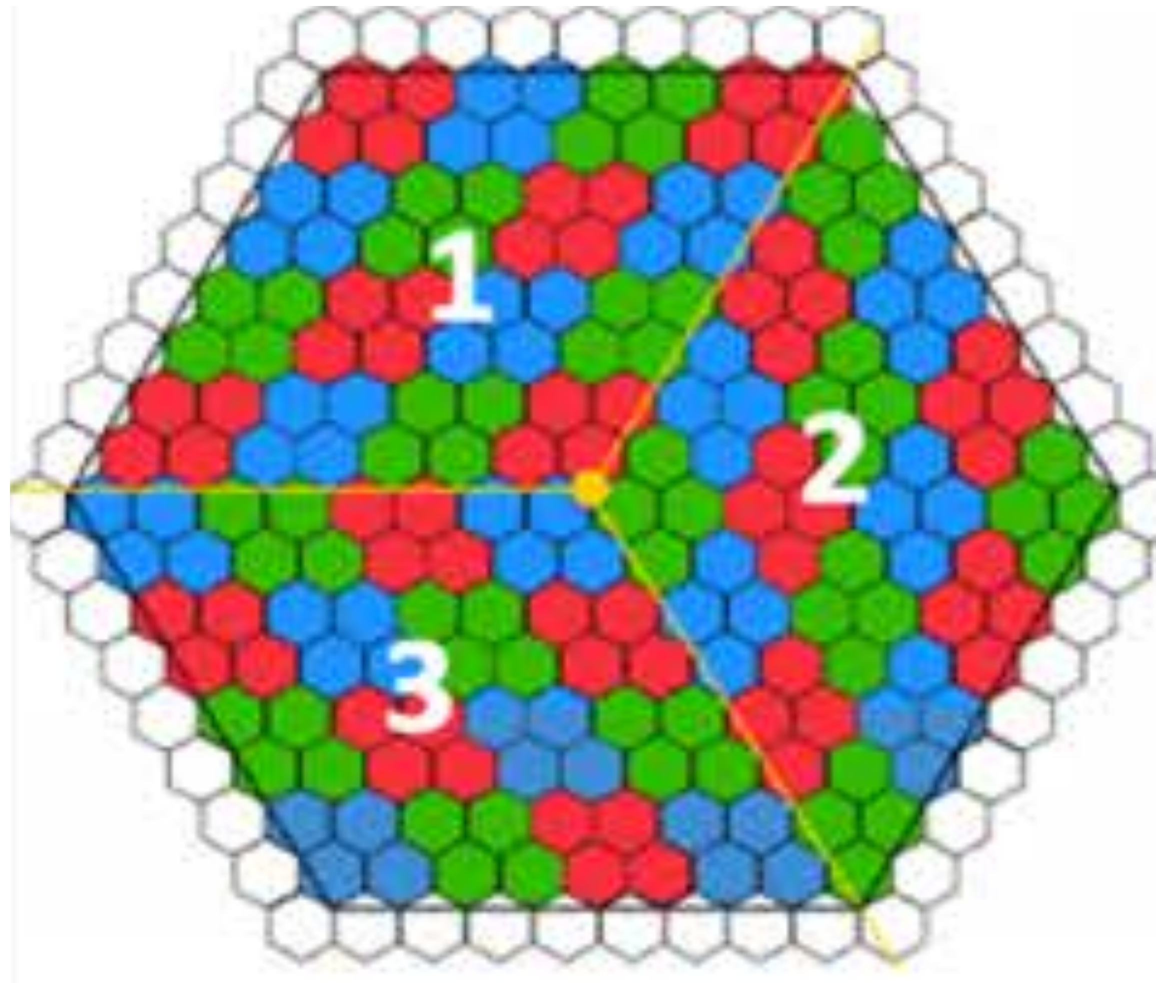
Encoded data



Transmit encoded data!

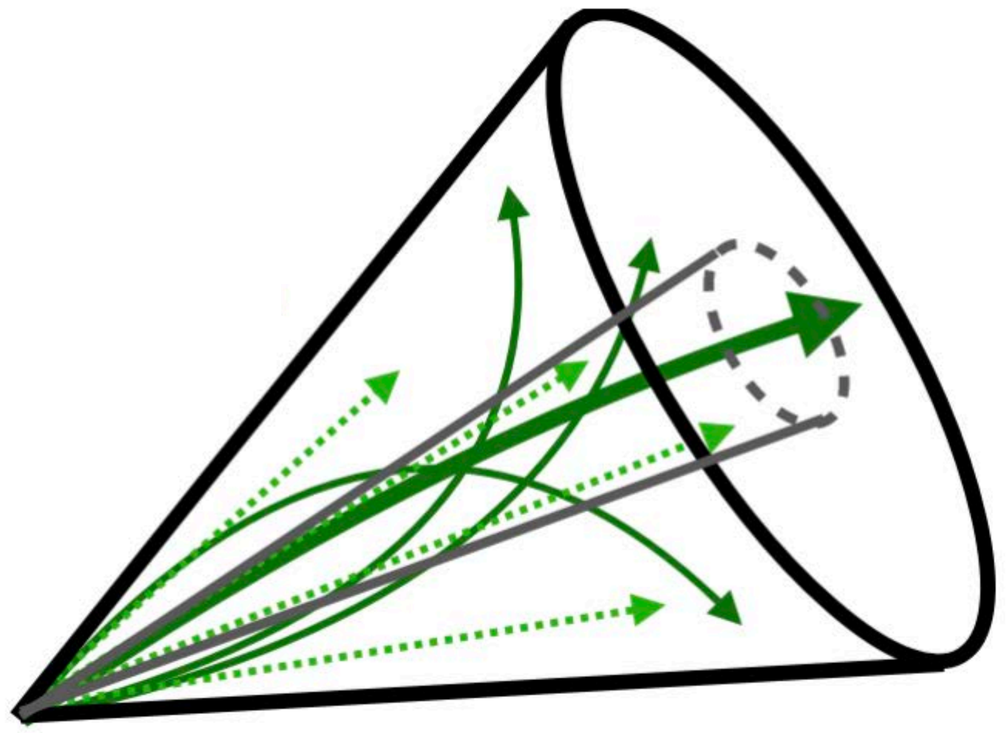


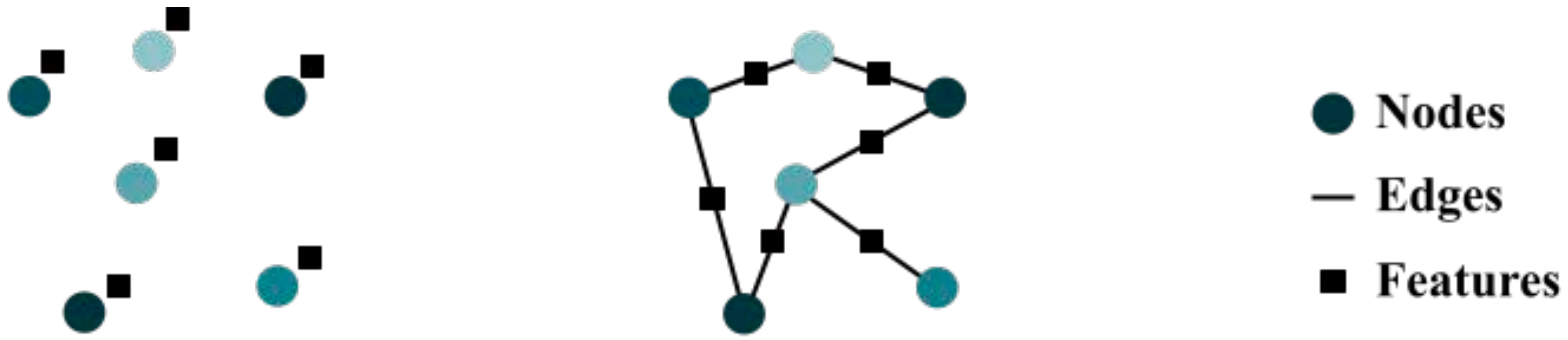
Encoded data



https://arxiv.org/abs/2402.01876

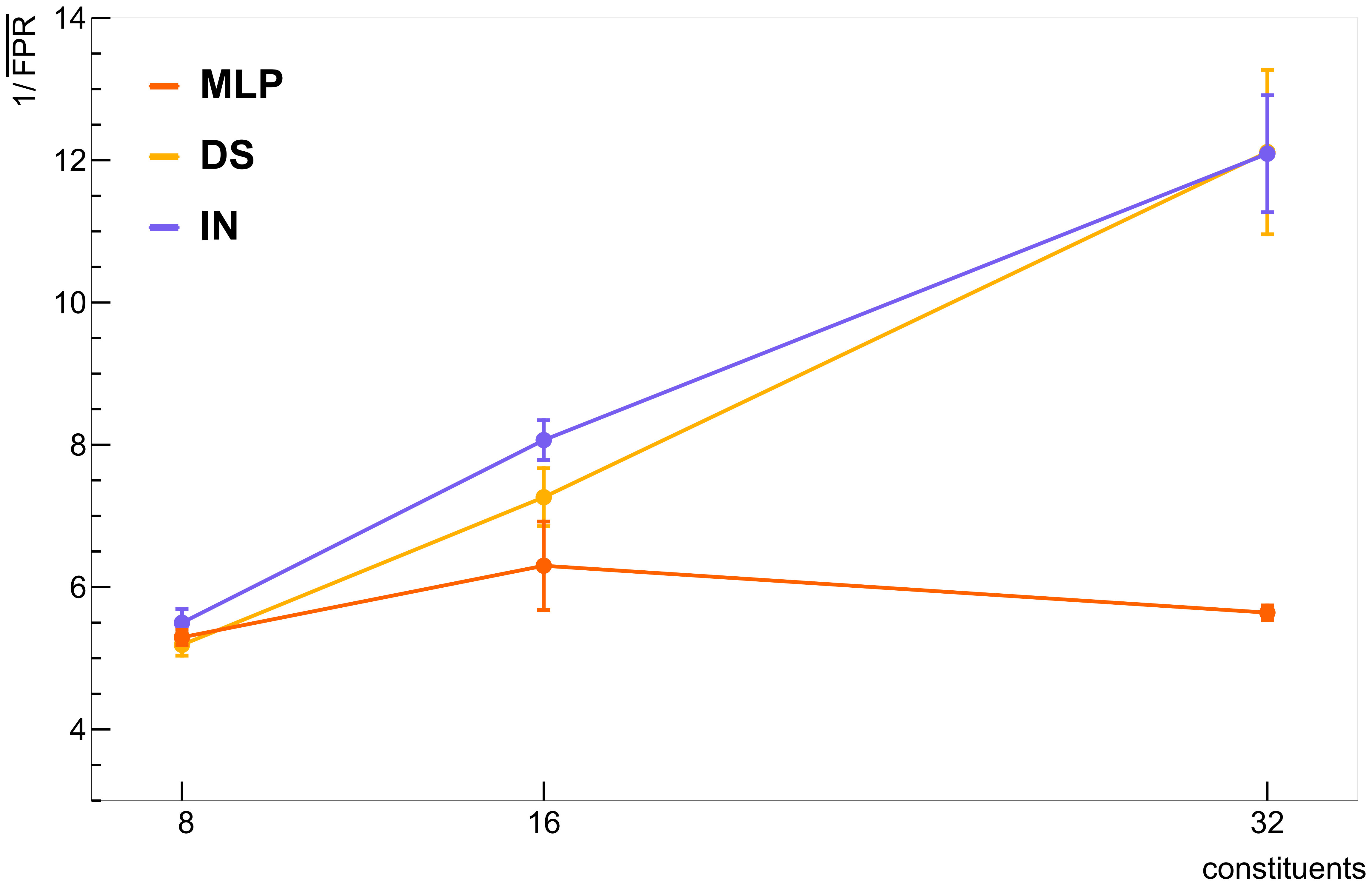
Invariance vs equivariance, sets vs graphs for smaller models?





Sets: Information is only assigned to individual nodes.

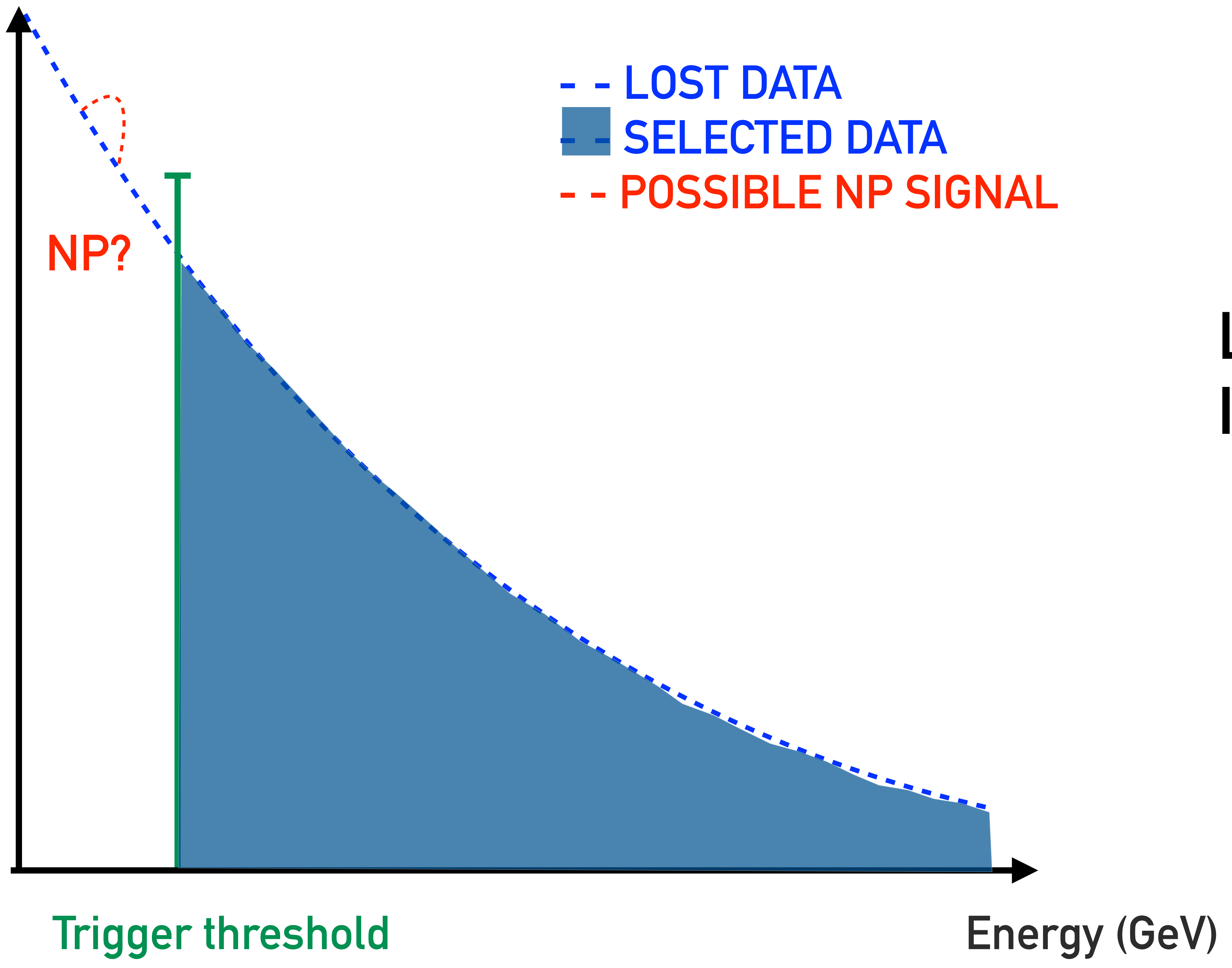
Graphs: Information is assigned to edges, i.e., pairs of nodes.



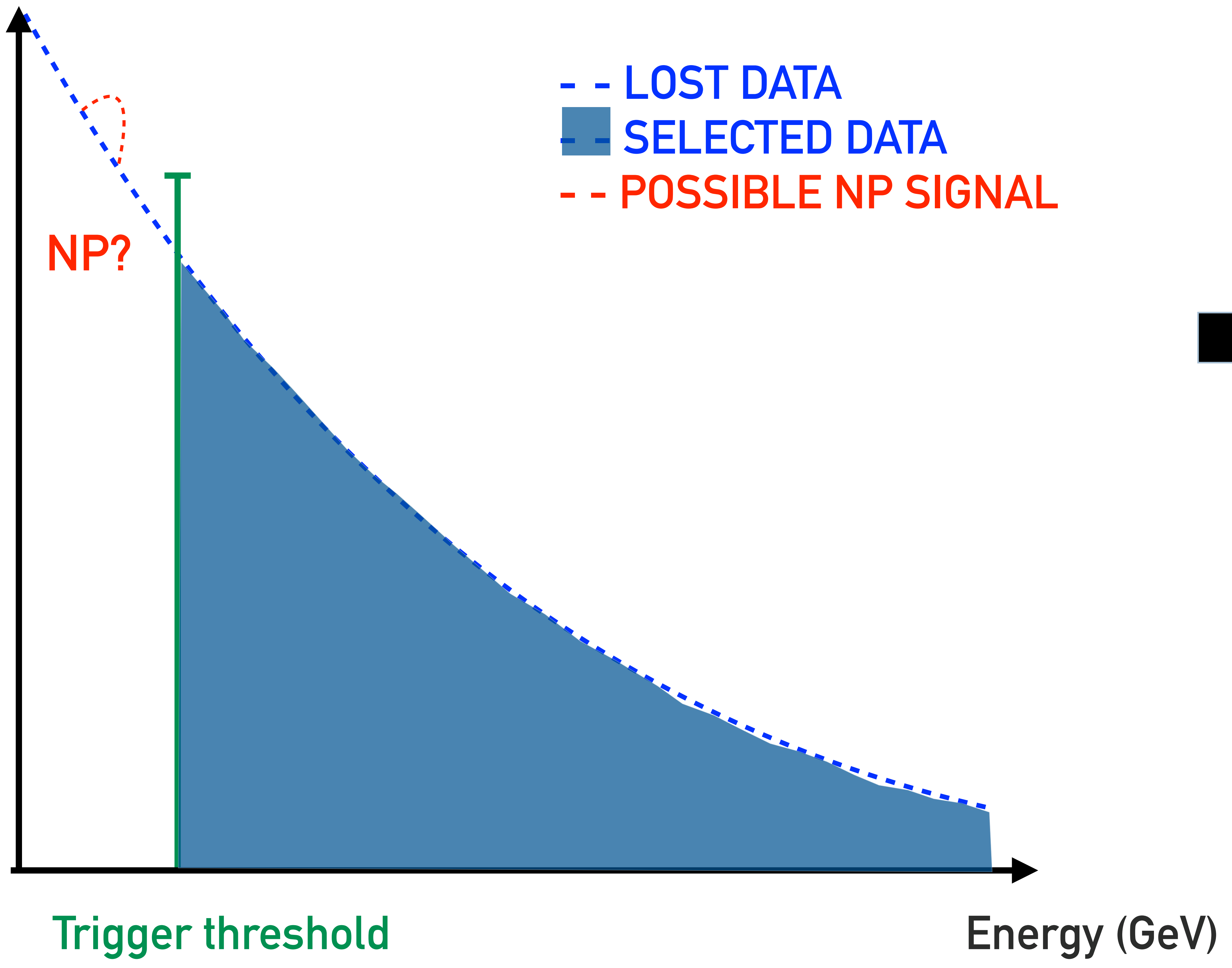
FPGA: Xilinx Virtex UltraScale+ VU13P

Architecture	Constituents	RF	Latency [ns] (cc)	II [ns] (cc)	DSP	LUT
MLP	8	1	105 (21)	5 (1)	262 (2.1%)	155,080 (9.0%)
	16	1	100 (20)	5 (1)	226 (1.8%)	146,515 (8.5%)
	32[a]	1	105 (21)	5 (1)	262 (2.1%)	155,080 (7.2%)
DS	8	2	95 (19)	15 (3)	626 (5.1%)	386,294 (22.3%)
	16	4	115 (23)	15 (3)	555 (4.5%)	747,374 (43.2%)
	32[a]	8	130 (26)	10 (2)	434 (3.5%)	903,284 (52.3%)
IN	8	2	160 (32)	15 (3)	2,191 (17.8%)	472,140 (27.3%)
	16	4	180 (36)	15 (3)	5,362 (43.6%)	1,387,923 (80.3%)
	32[a]	8	205 (41)	15 (3)	2,120 (17.3%)	1,162,104 (67.3%)

Limitations of current trigger

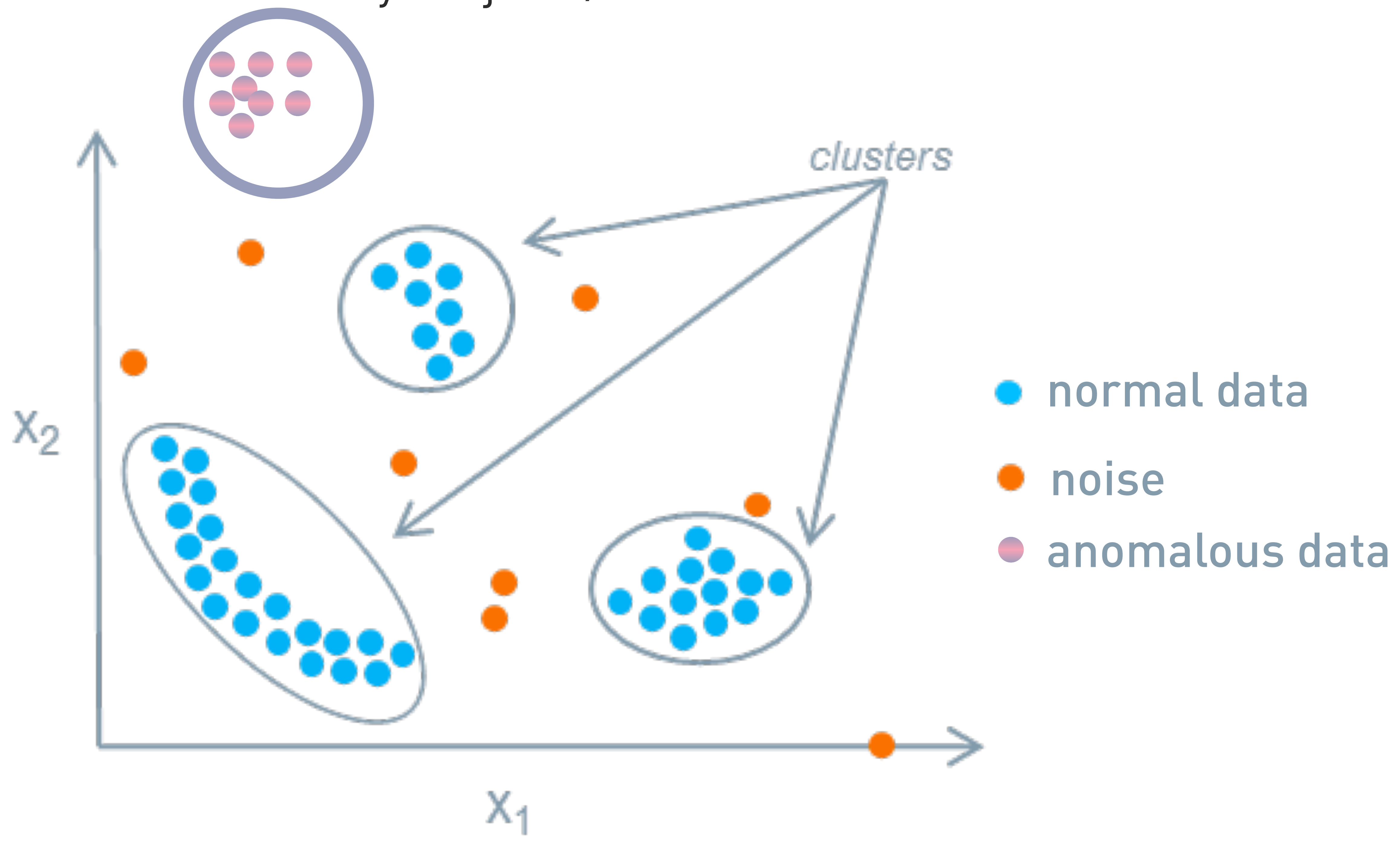


Level-1 rejects >99% of events!
Is there a smarter way to select?



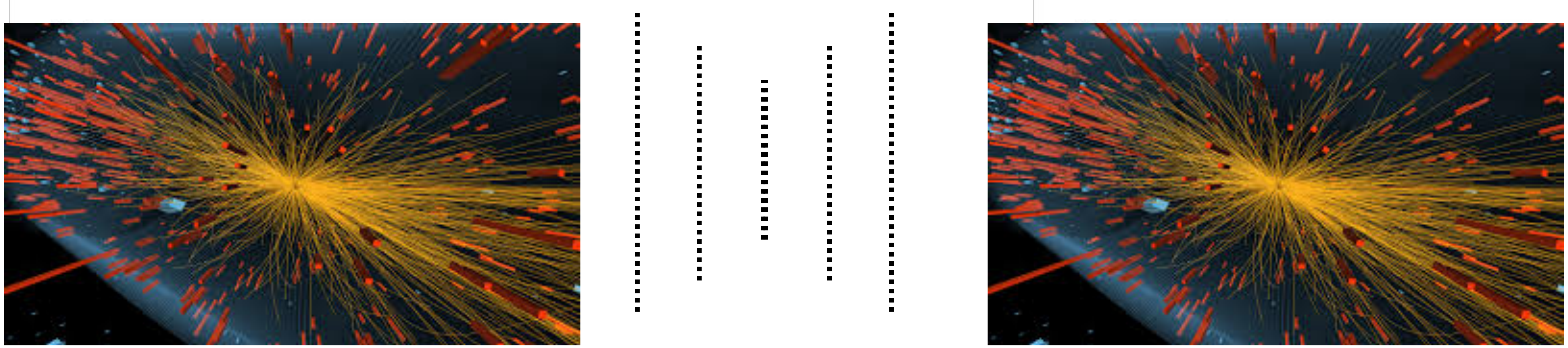
Look at **data** rather than defining signal hypothesis a priori

- Can we “classify” objects/events?



AXOL1TL

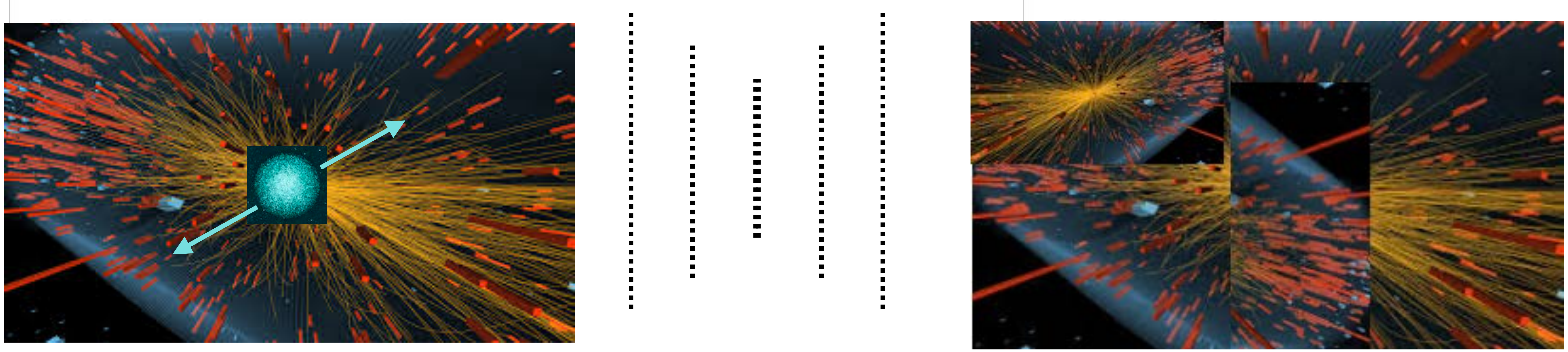
$\mathbf{x}$ $\hat{\mathbf{x}}$



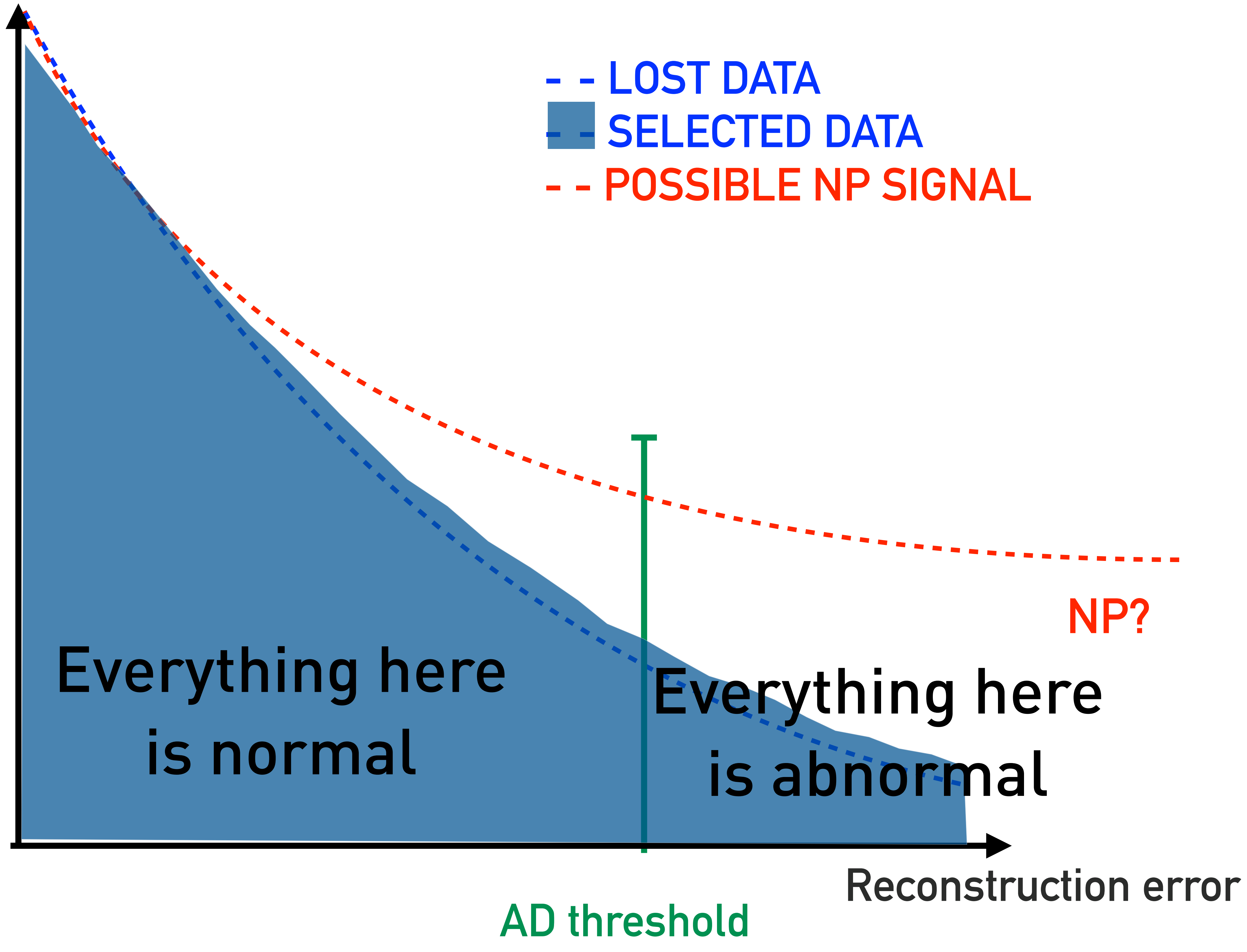
$$\text{loss} = ||\,x - \hat{x}\,||^2$$

AXOL1TL

$\mathbf{x}$ $\hat{\mathbf{x}}$



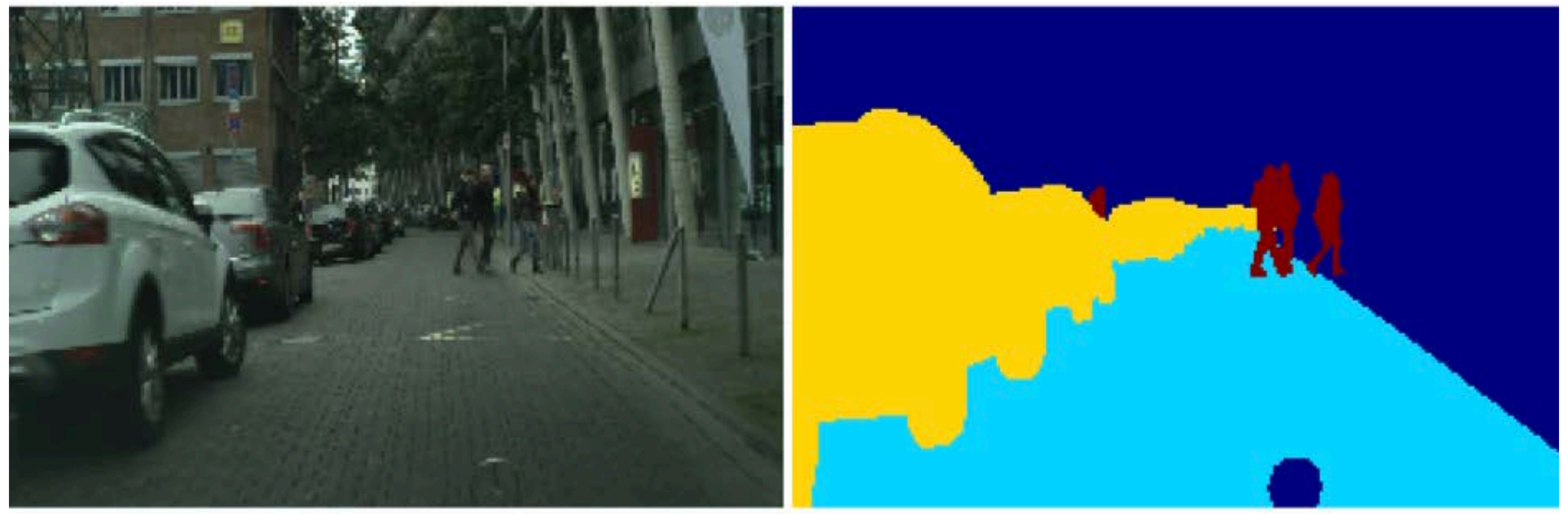
$$\text{loss} = ||\mathbf{x} - \hat{\mathbf{x}}||^2$$



AXOL1TL

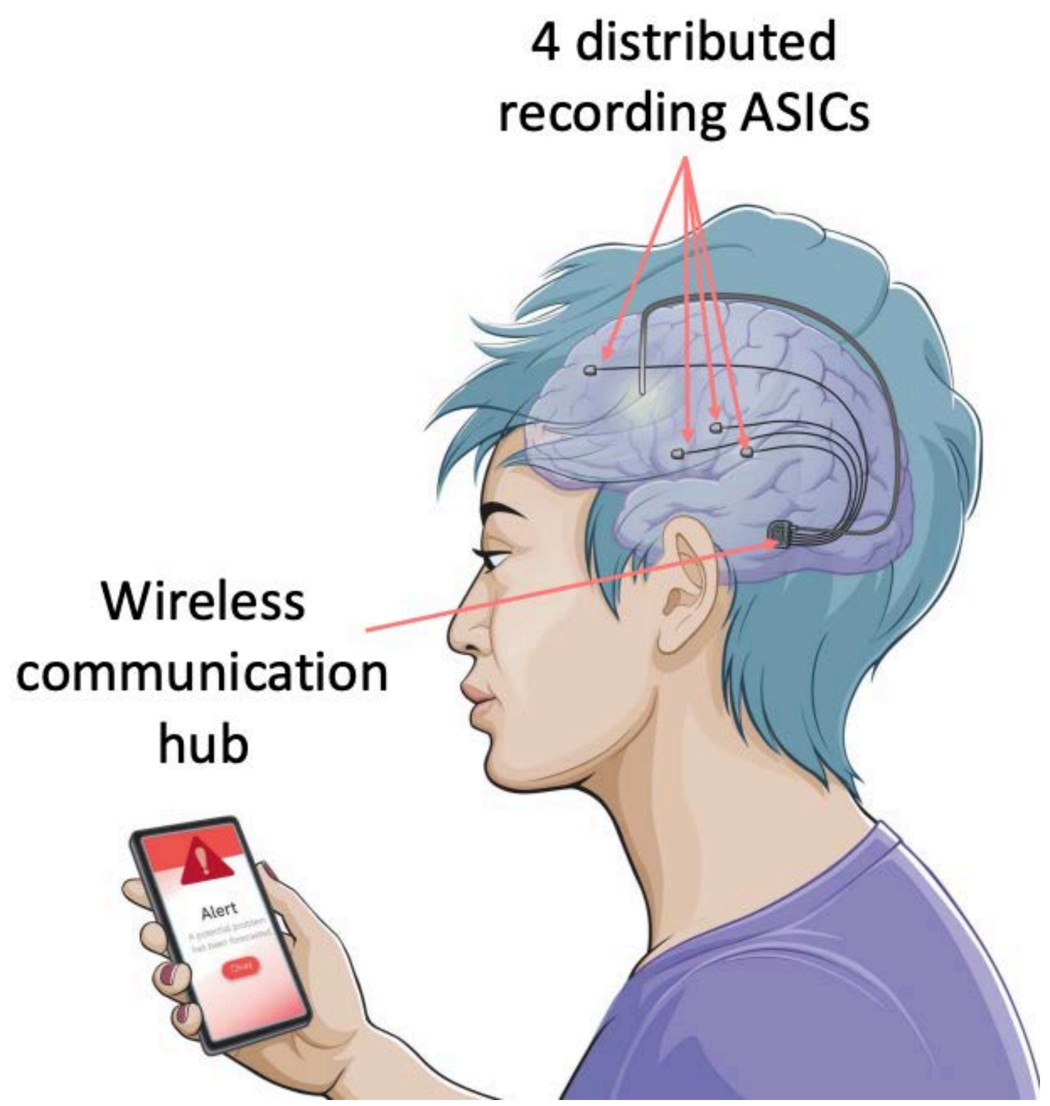
....in 50 nanoseconds!

Semantic segmentation for autonomous vehicles



N. Ghielmetti et al.

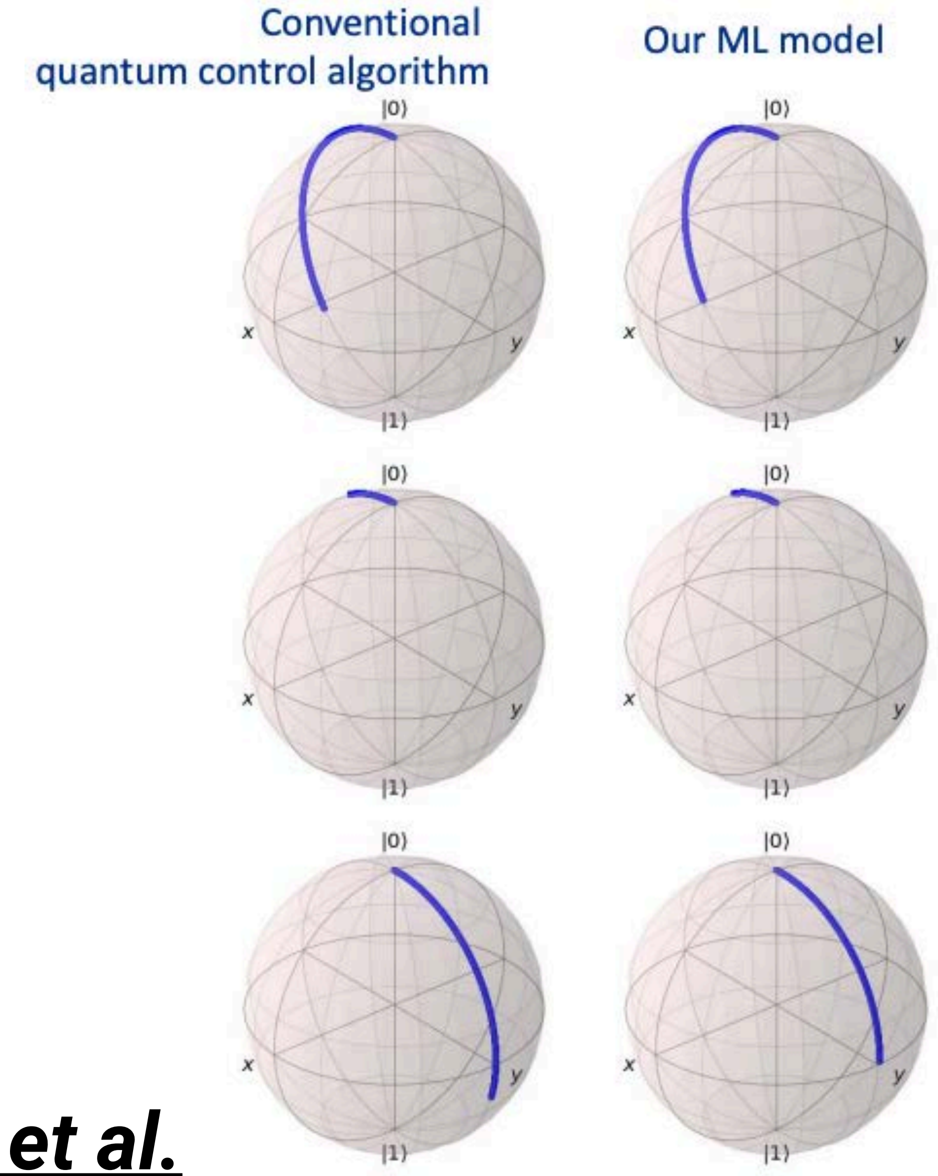
Seizure Predicting Brain Implant



W. Lemaire et al.

NN accelerator for quantum control

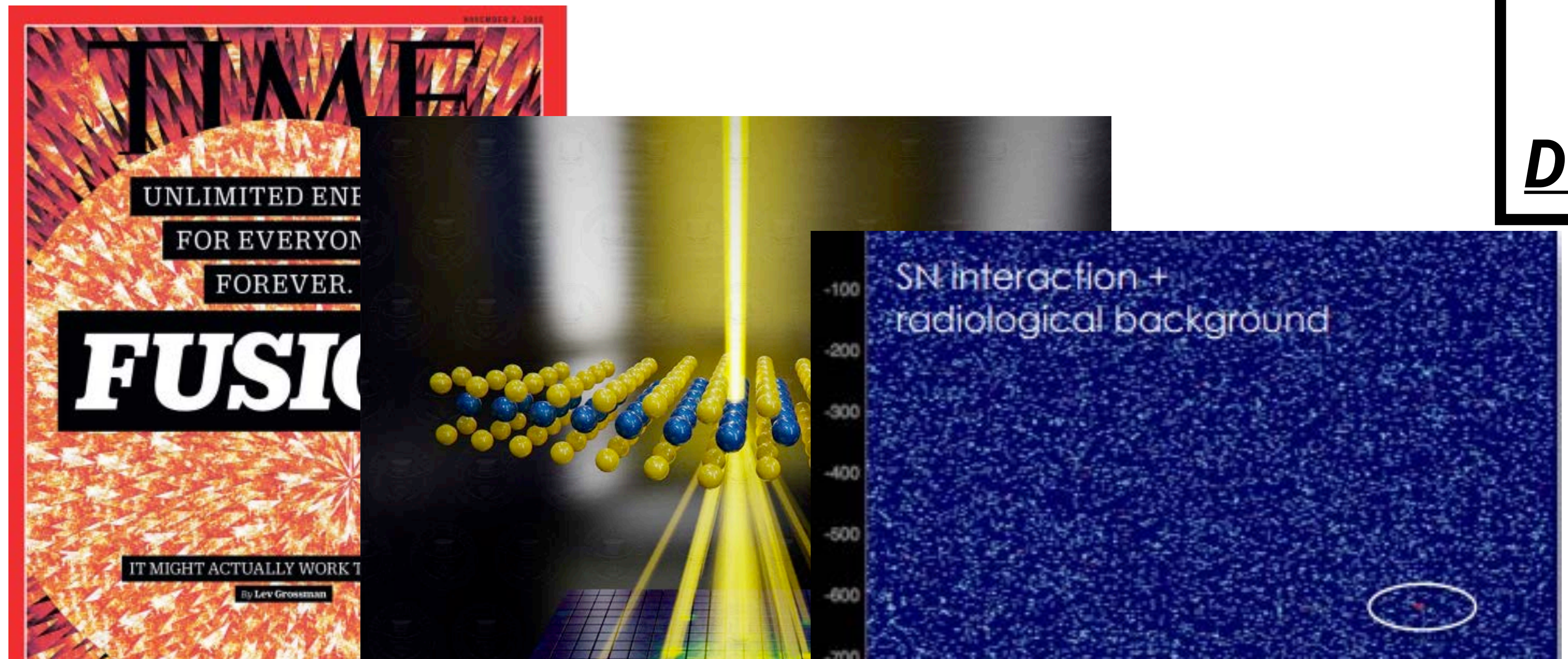
- Putting control in cryostat (e.g optimal pulse parameters)



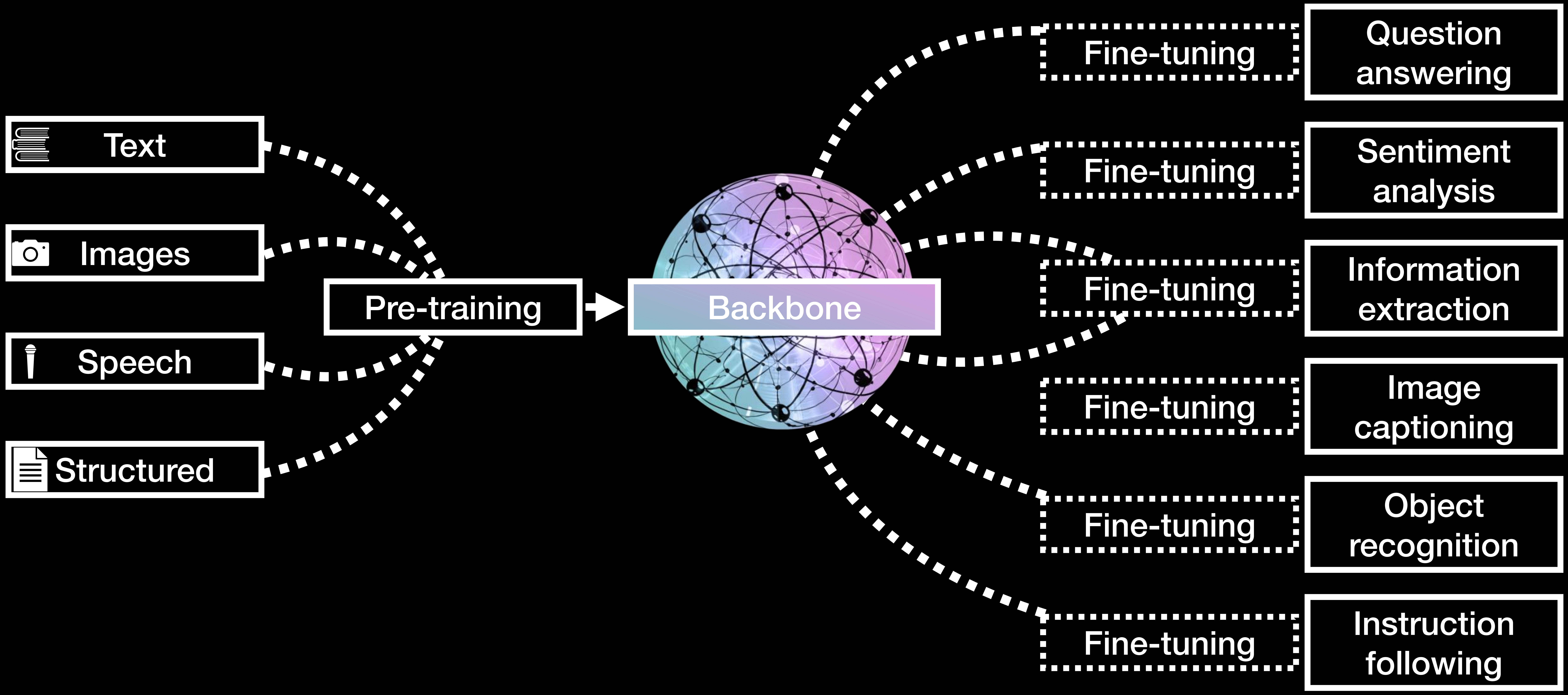
D Xu et al.

Other examples

- For fusion science phase/mode monitoring
- Crystal structure detection
- Triggering in DUNE
- Accelerator control
- Magnet Quench Detection
- MLPerf tinyML benchmarking
- Food contamination detection
- etc….

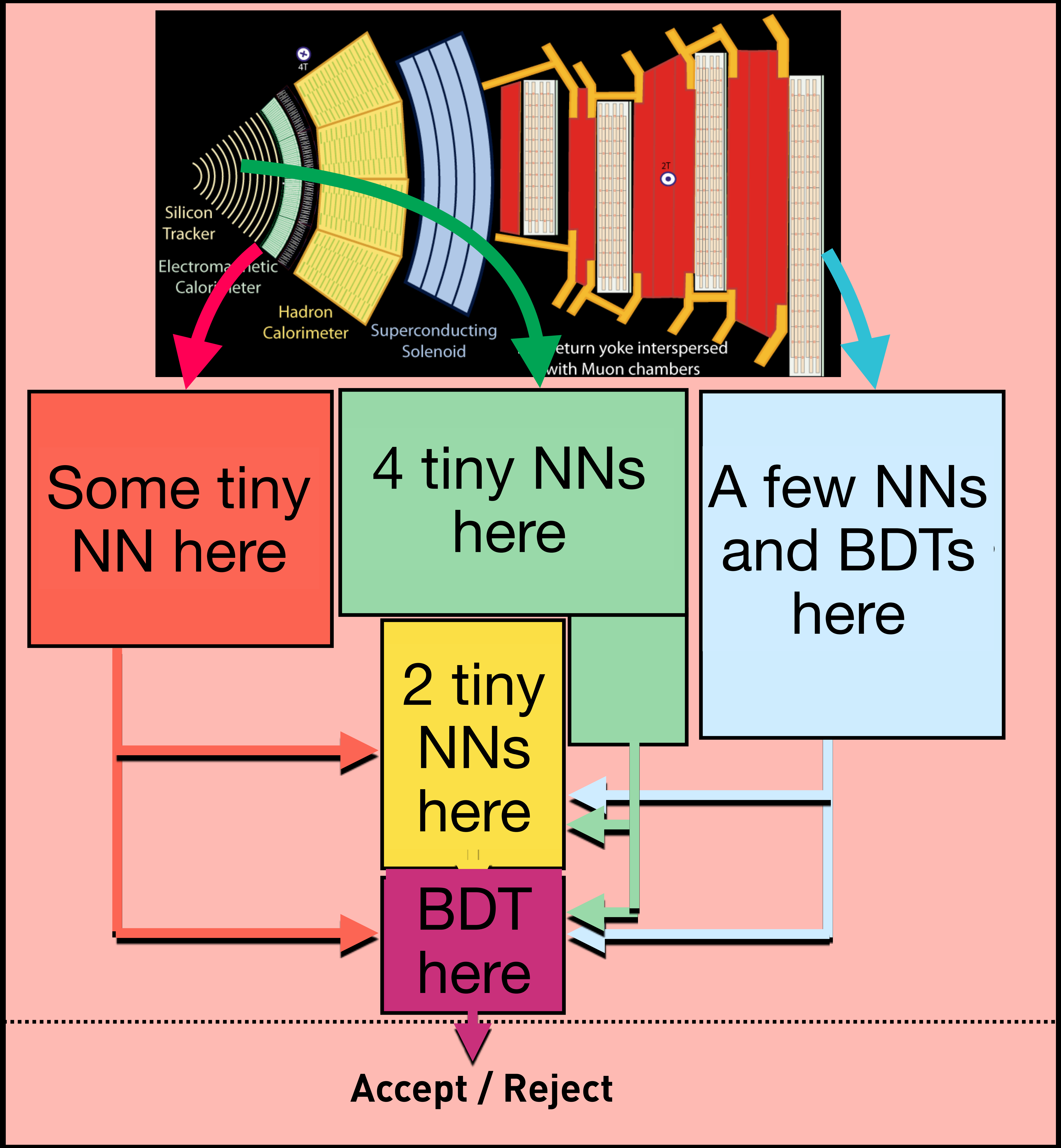


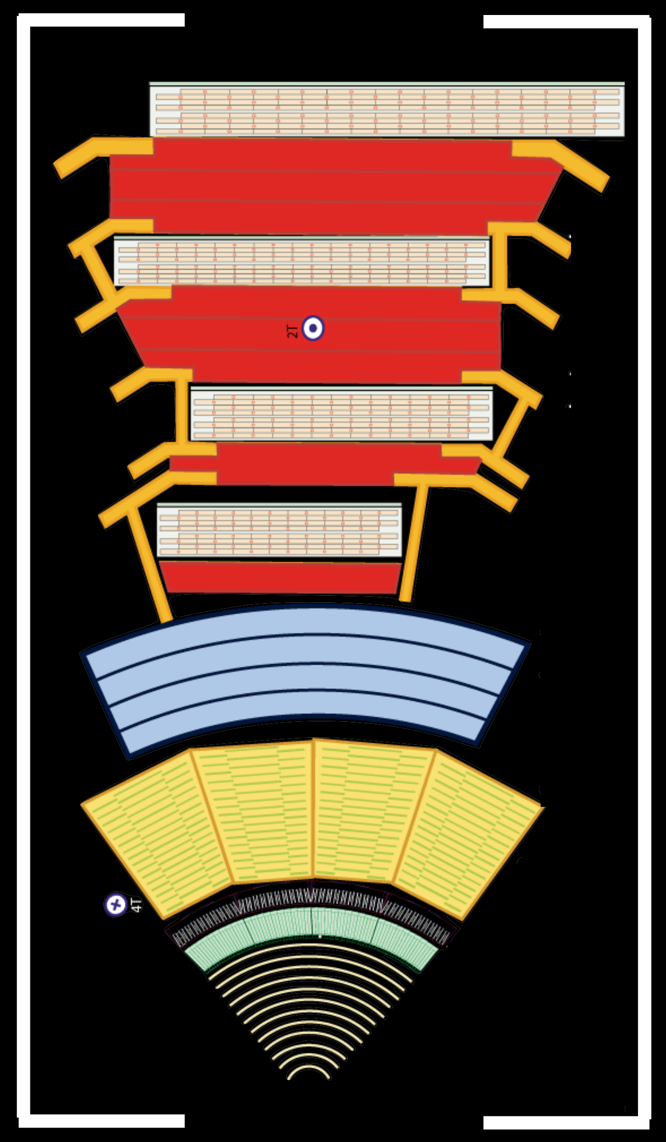
63 Tb/s
Silicon Tracker
Electromagnetic Calorimeter
Hadron Calorimeter
Superconducting Solenoid
Iron return yoke interspersed with Muon chambers
CALORIMETRY (370 FPGAs)
CHARGED PARTICLE TRACKING (174 FPGAs)
MUON CHAMBERS (96 FPGAs)
PARTICLE FLOW (66 FPGAs)
GLOBAL TRIGGER (13 FPGAs)
Accept / Reject
Current HL-LHC design
CALORIMETER PRE-PROCESSING
TRACKING PRE-PROCESSING
MUON CHAMBERS PRE-PROCESSING
End-to-end reconstruction model
Latent representation
Downstream Task
Downstream Task
Downstream Task
Downstream Task
12.5 μs
Accept / Reject
This project



Text
Images
Speech
Structured
Pre-training
Backbone
Fine-tuning
Question answering
Fine-tuning
Sentiment analysis
Fine-tuning
Information extraction
Fine-tuning
Image captioning
Fine-tuning
Object recognition
Fine-tuning
Instruction following


Heterogeneous detector
Multi-modal input!
Pre-training
Backbone
Fine-tuning
Jet reconstruction
Fine-tuning
Electron reconstruction
Fine-tuning
Pile-up removal
Fine-tuning
Missing energy computation
Fine-tuning
Anomaly Detection
Fine-tuning
0/1?
(Generate simulation?)

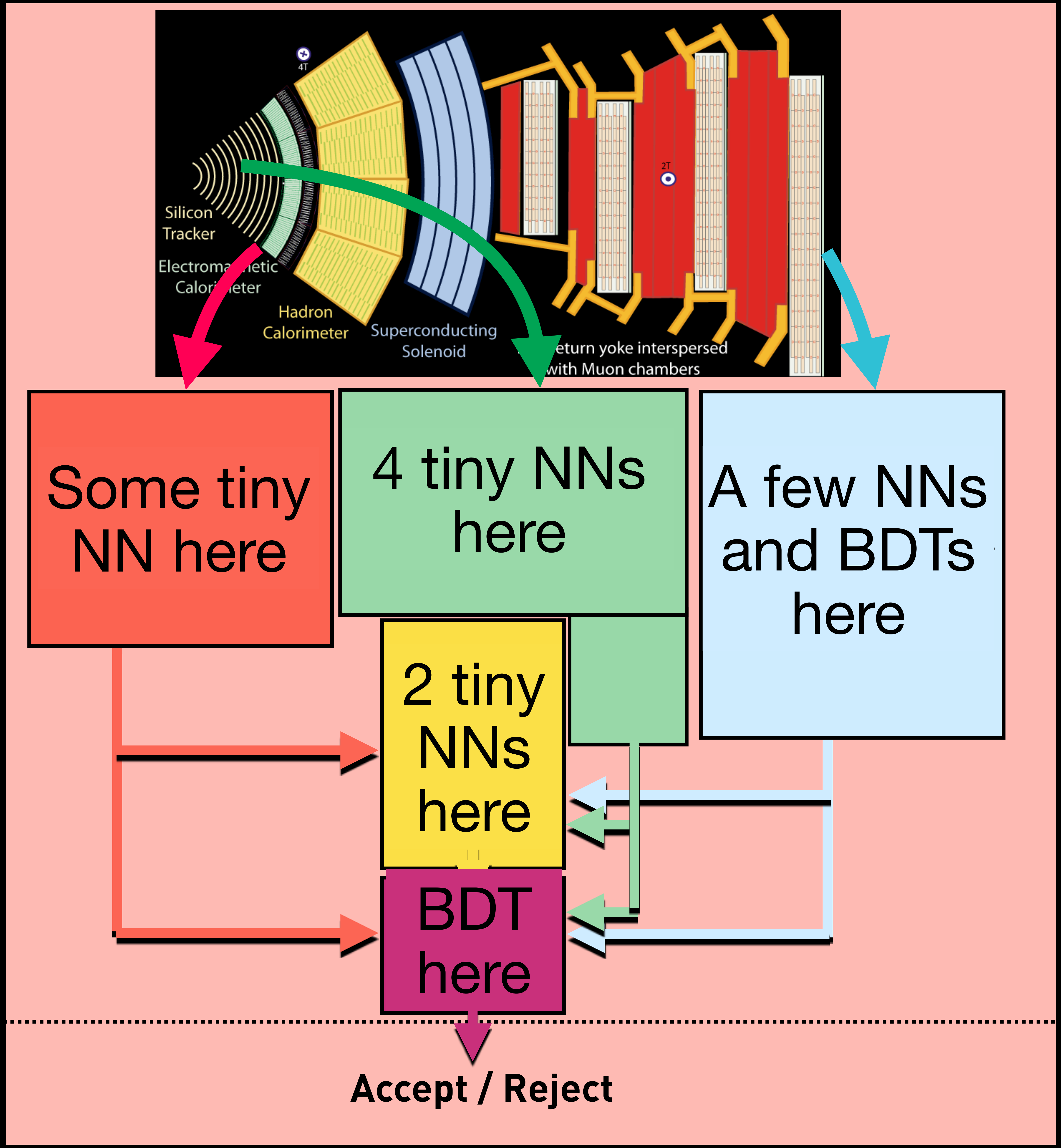


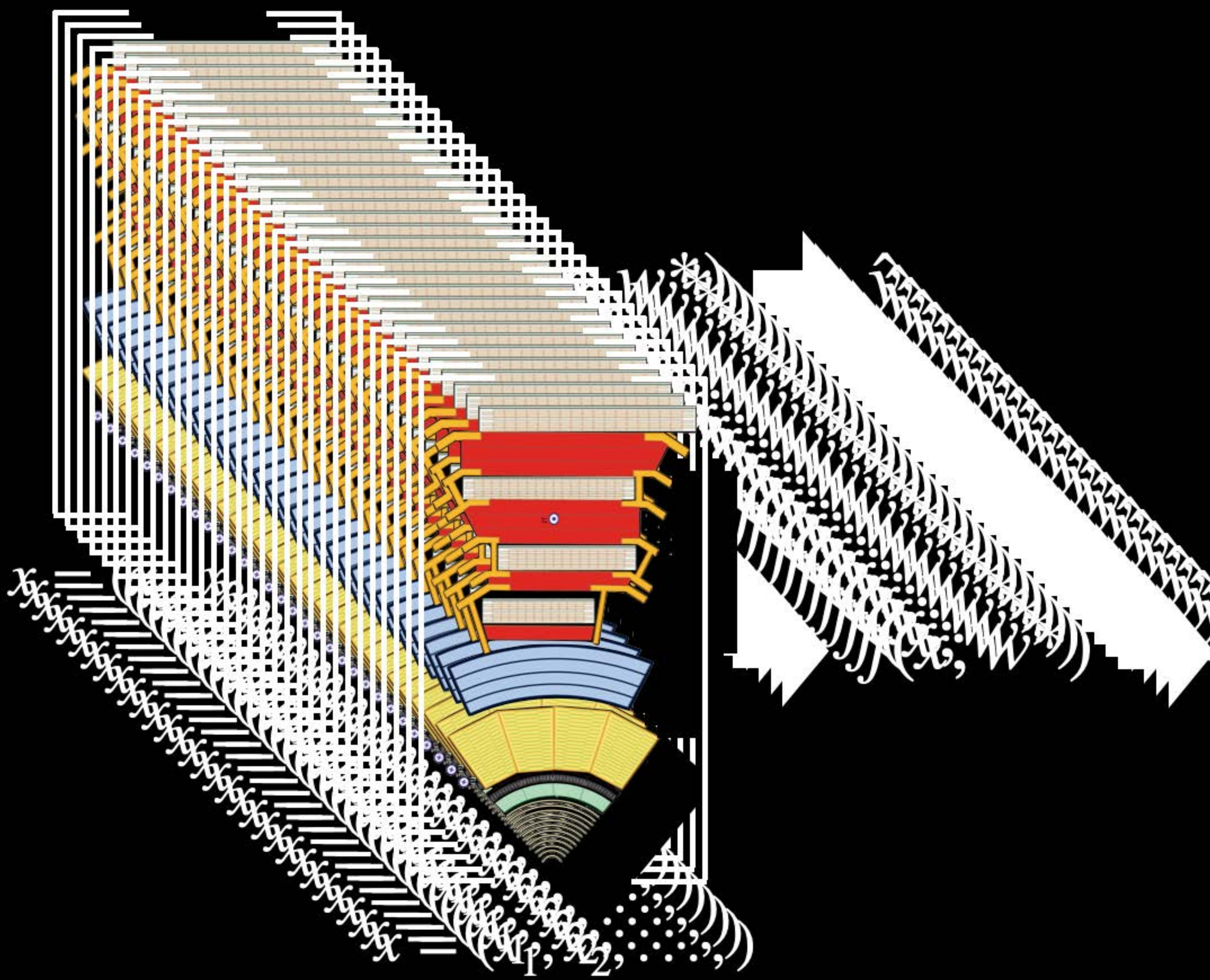


$$\Rightarrow f(x; w^*) \Rightarrow \hat{y}$$

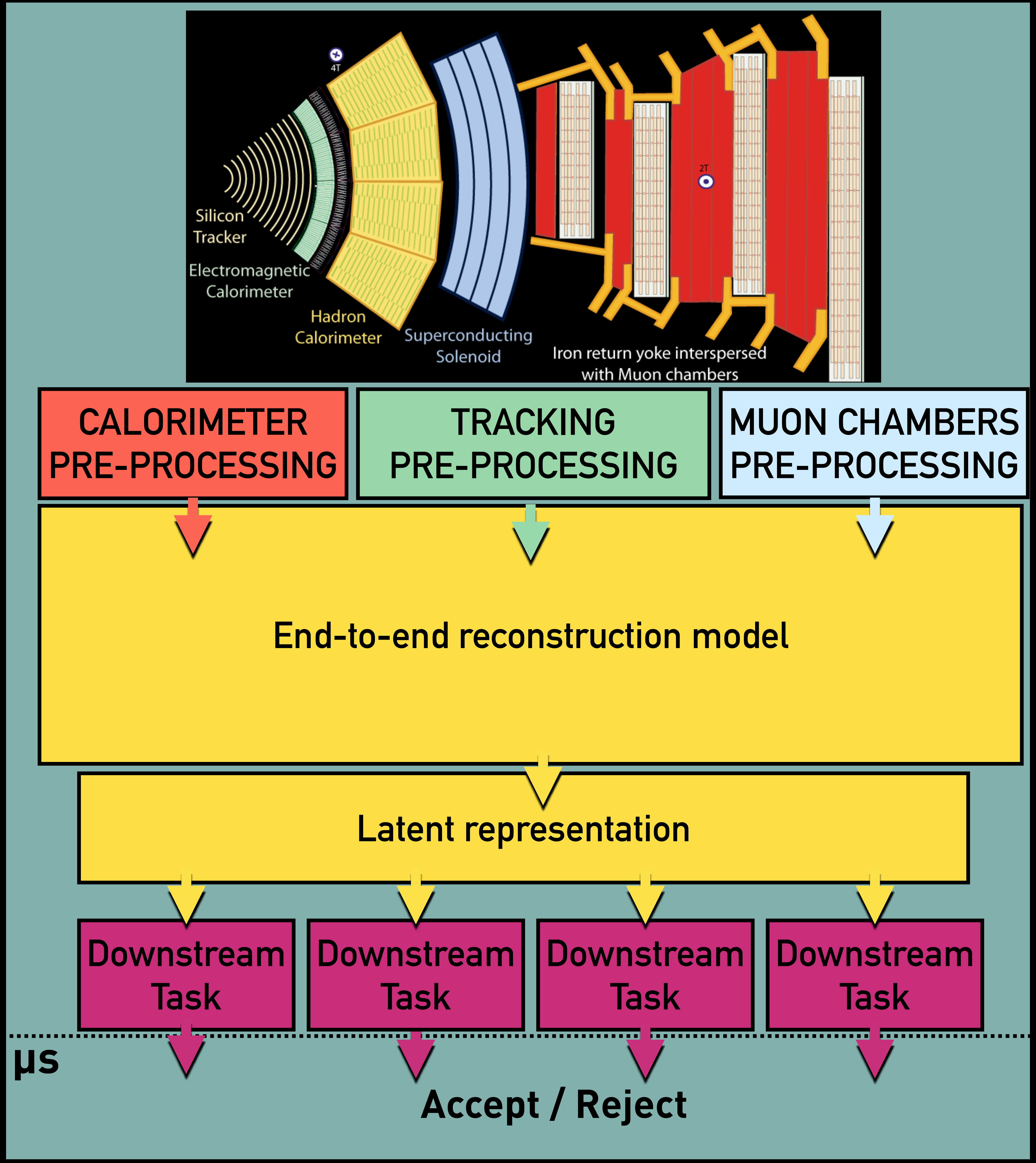
$$x = (x_1, x_2, \ldots,)$$

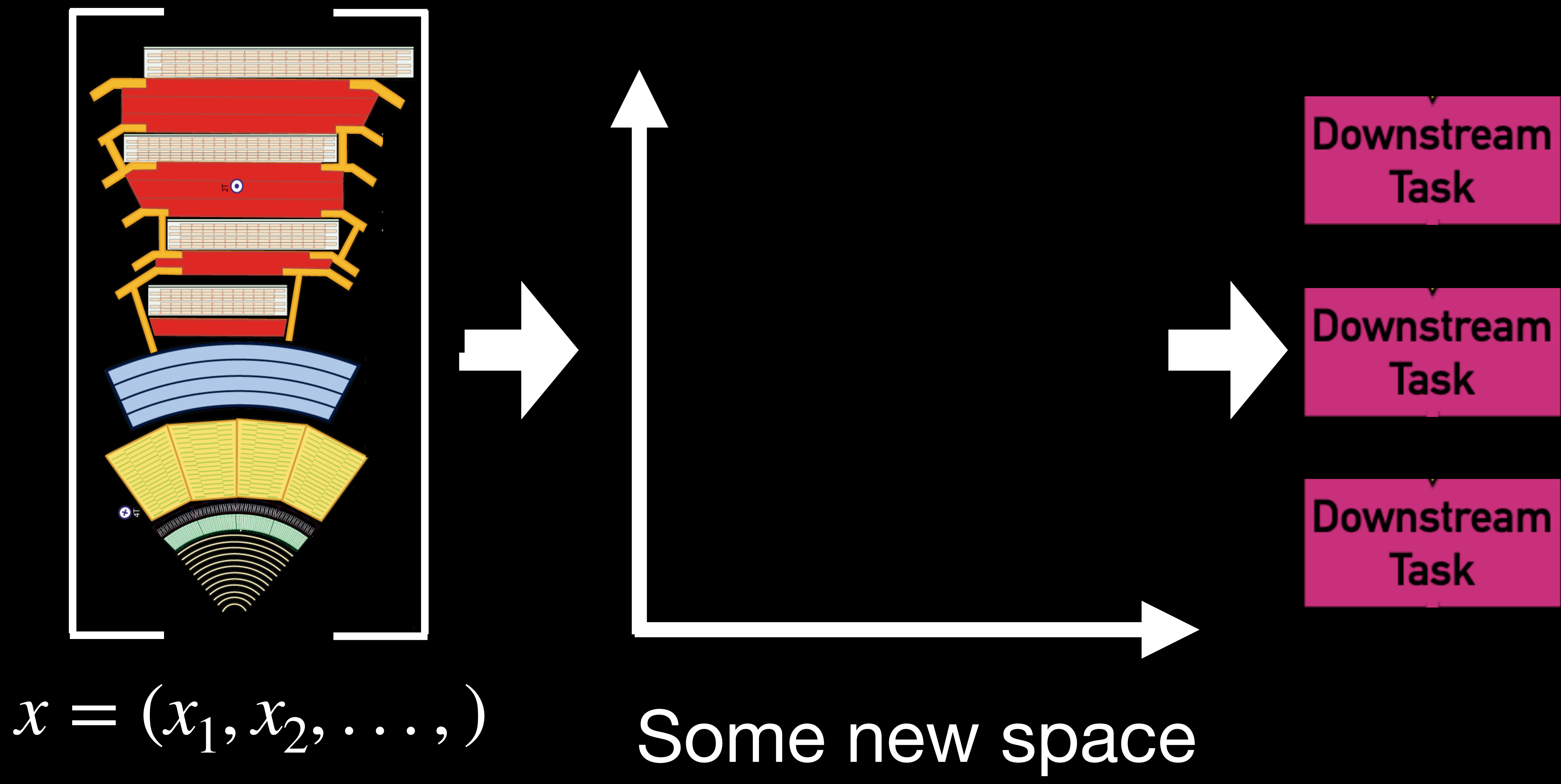
Too many models, too little learning?



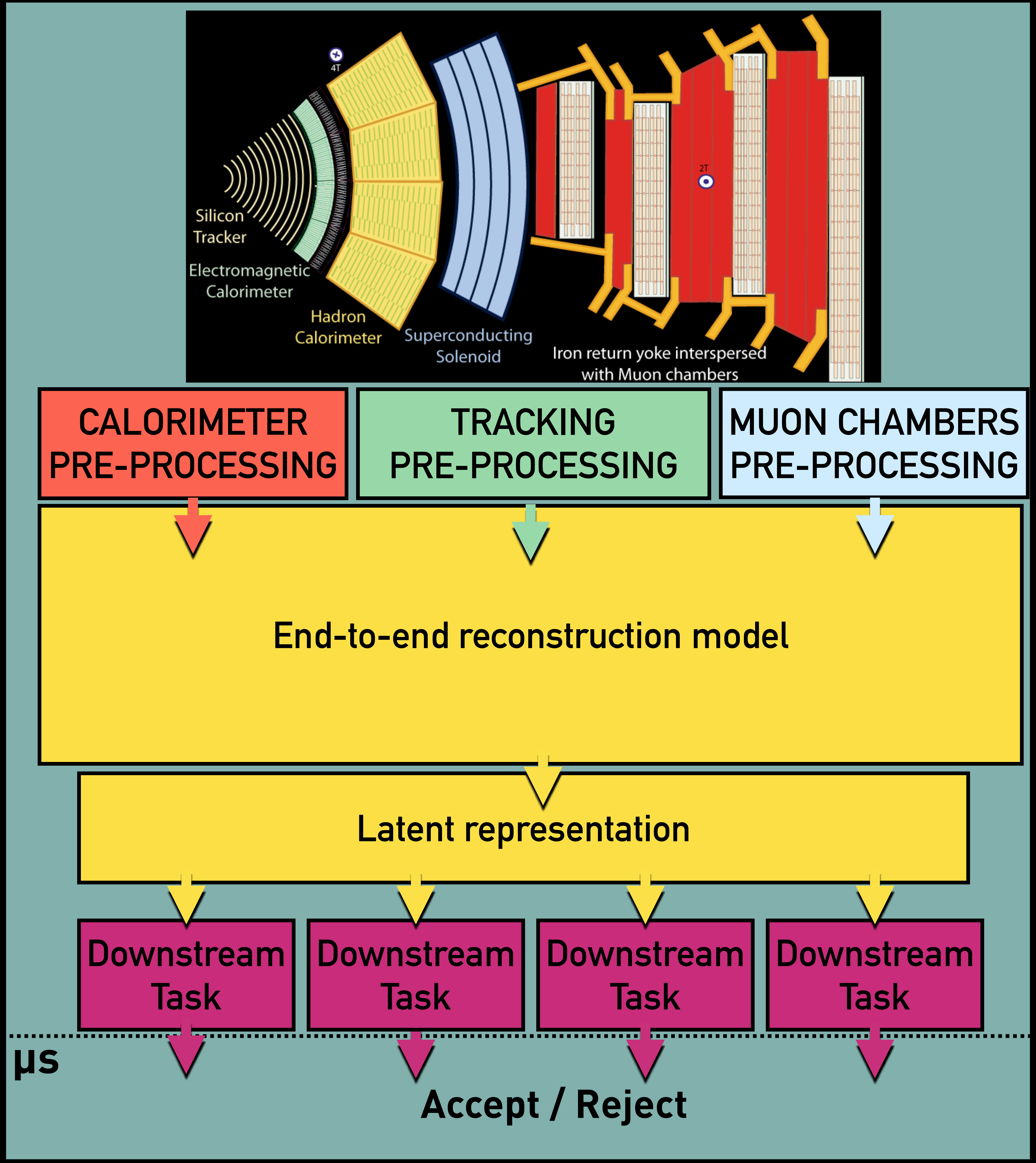


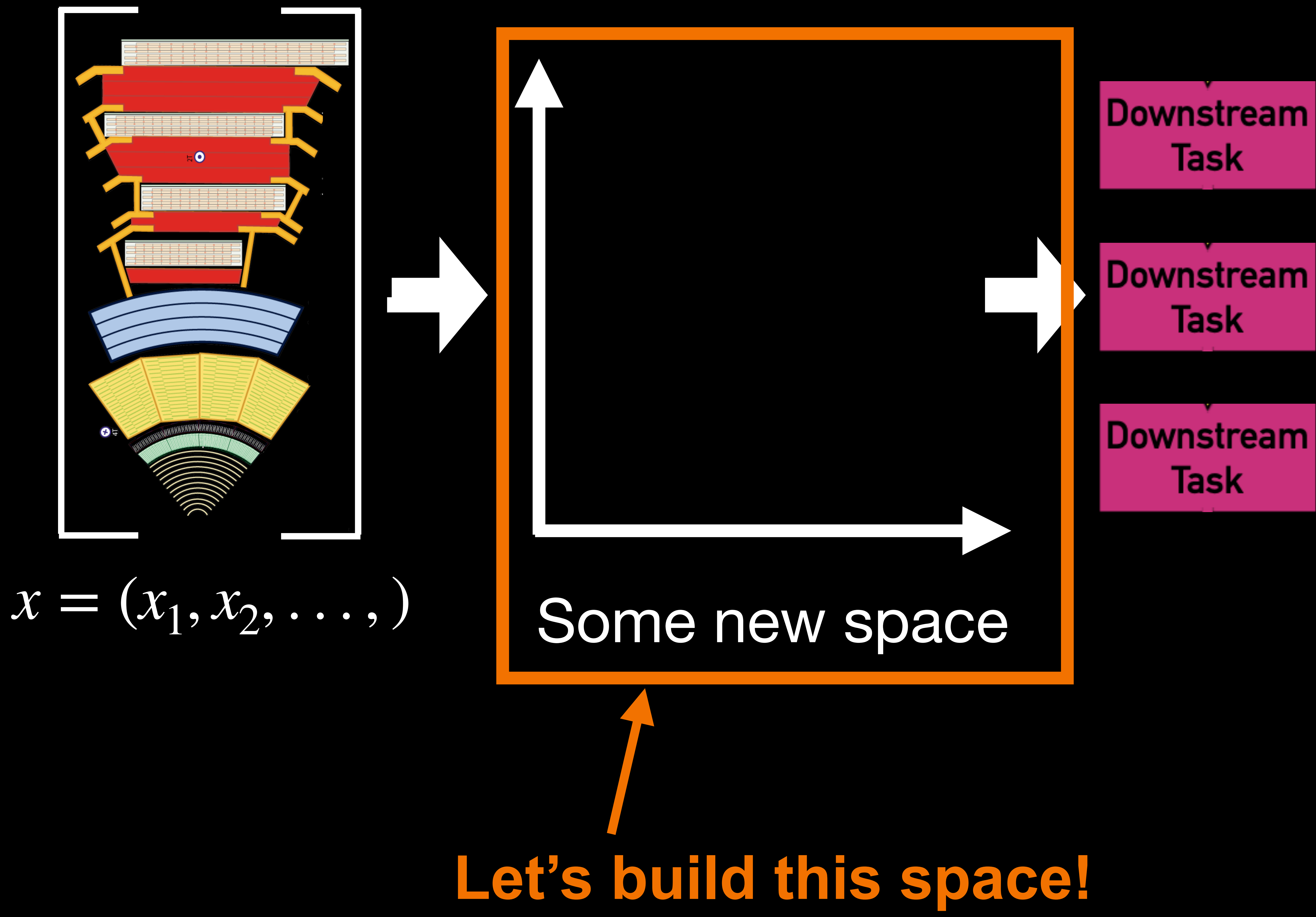
One model, learn neural embedding?



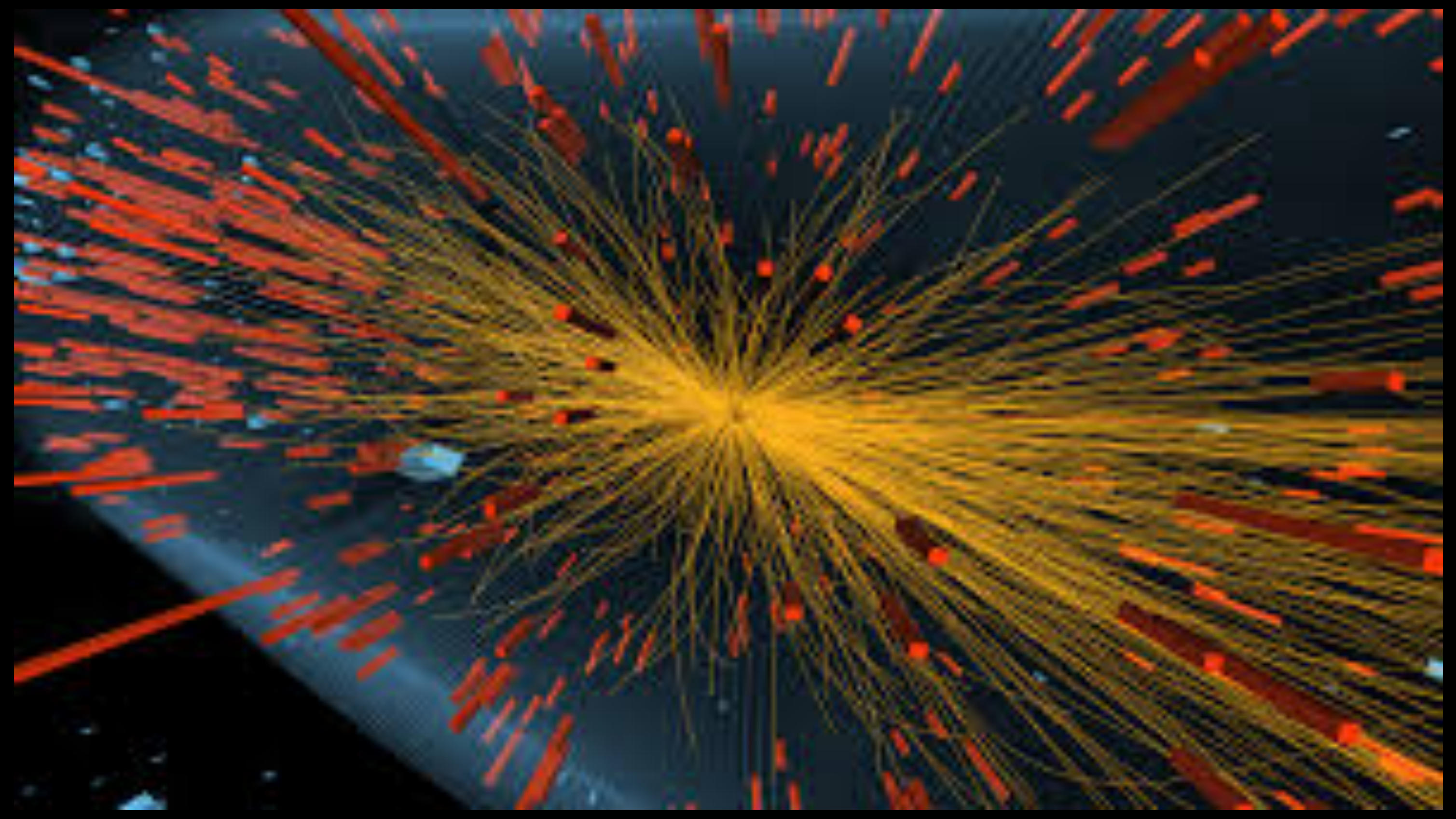


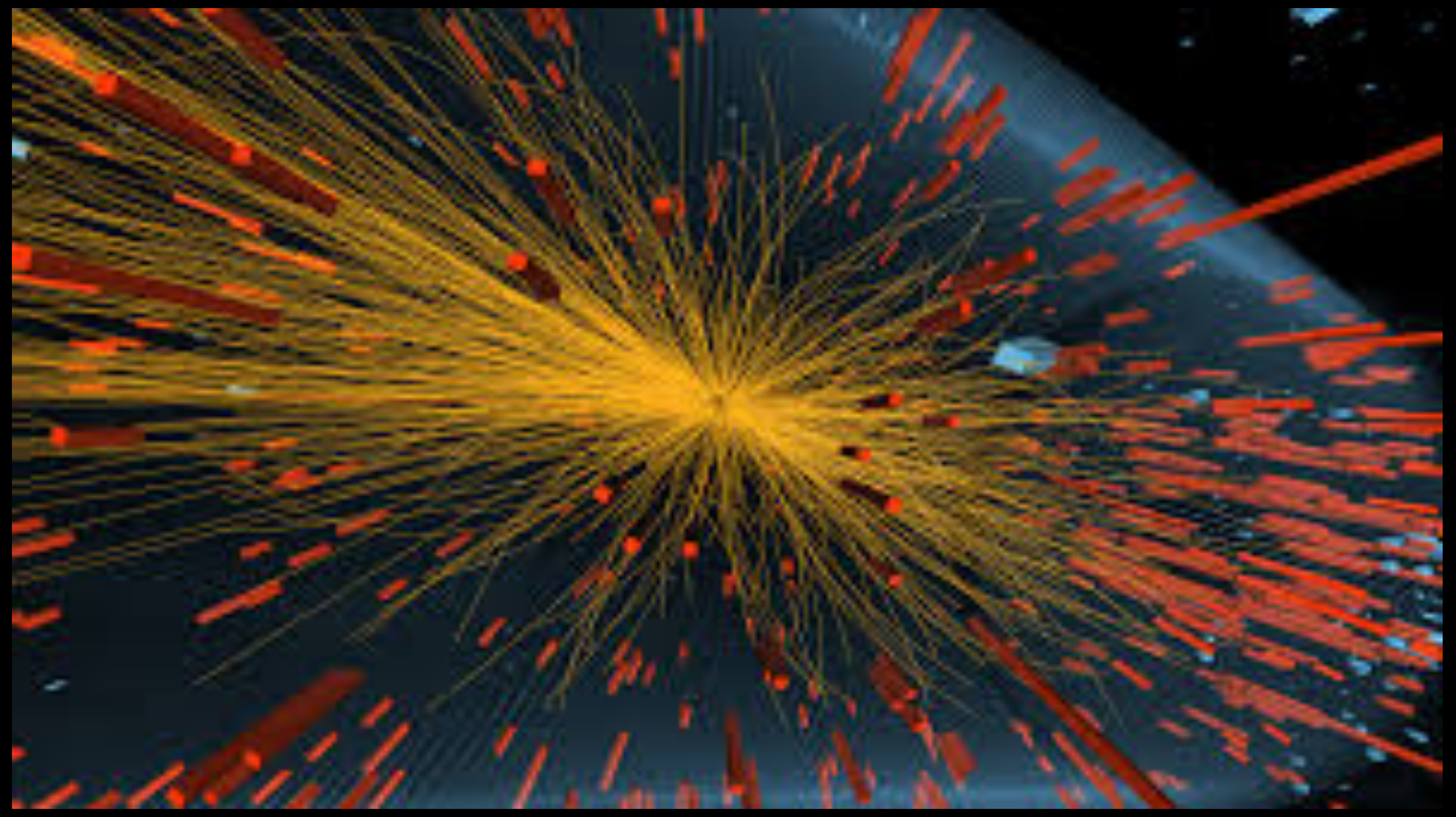
One model, learn neural embedding?





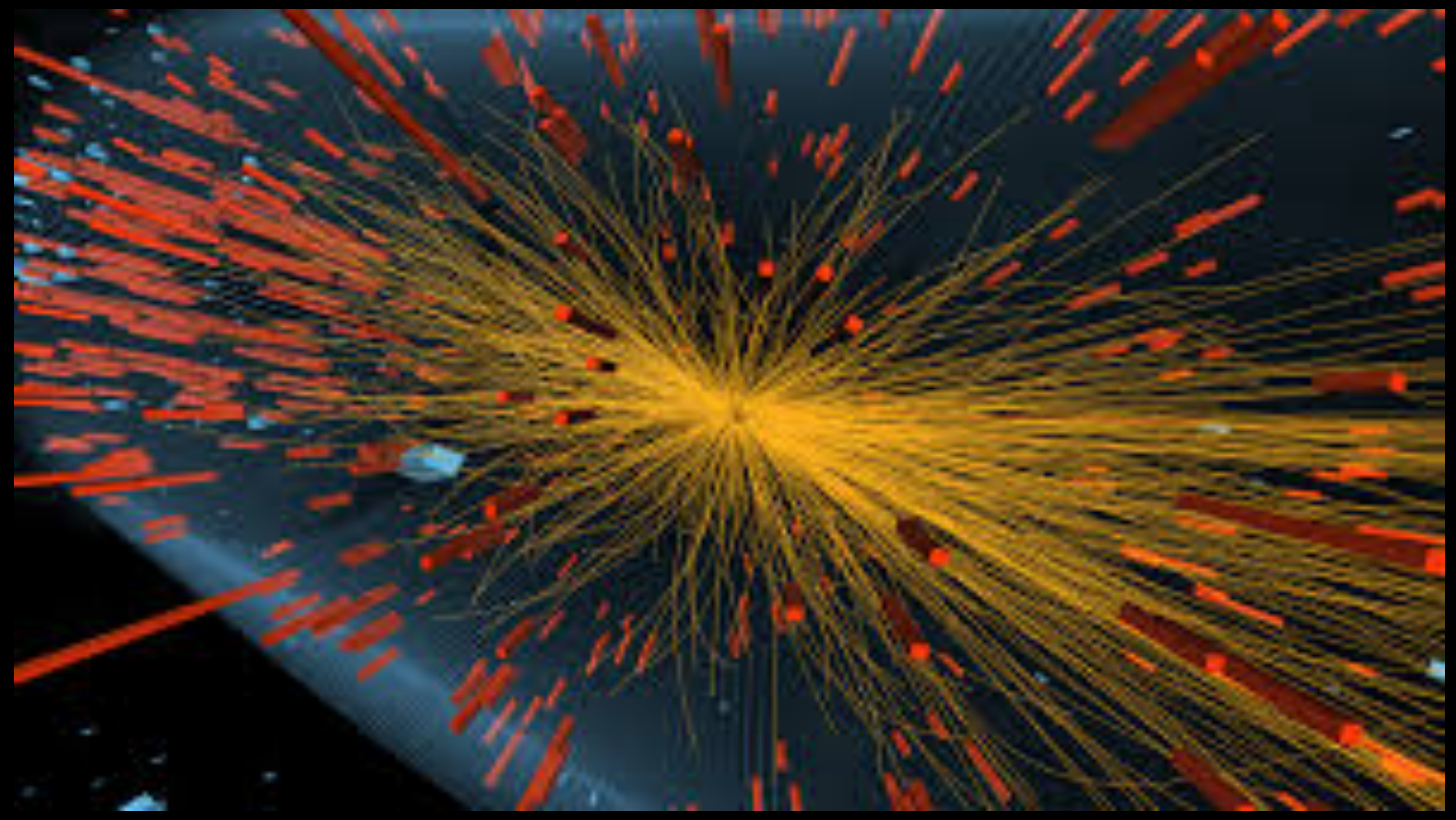
Learning the space

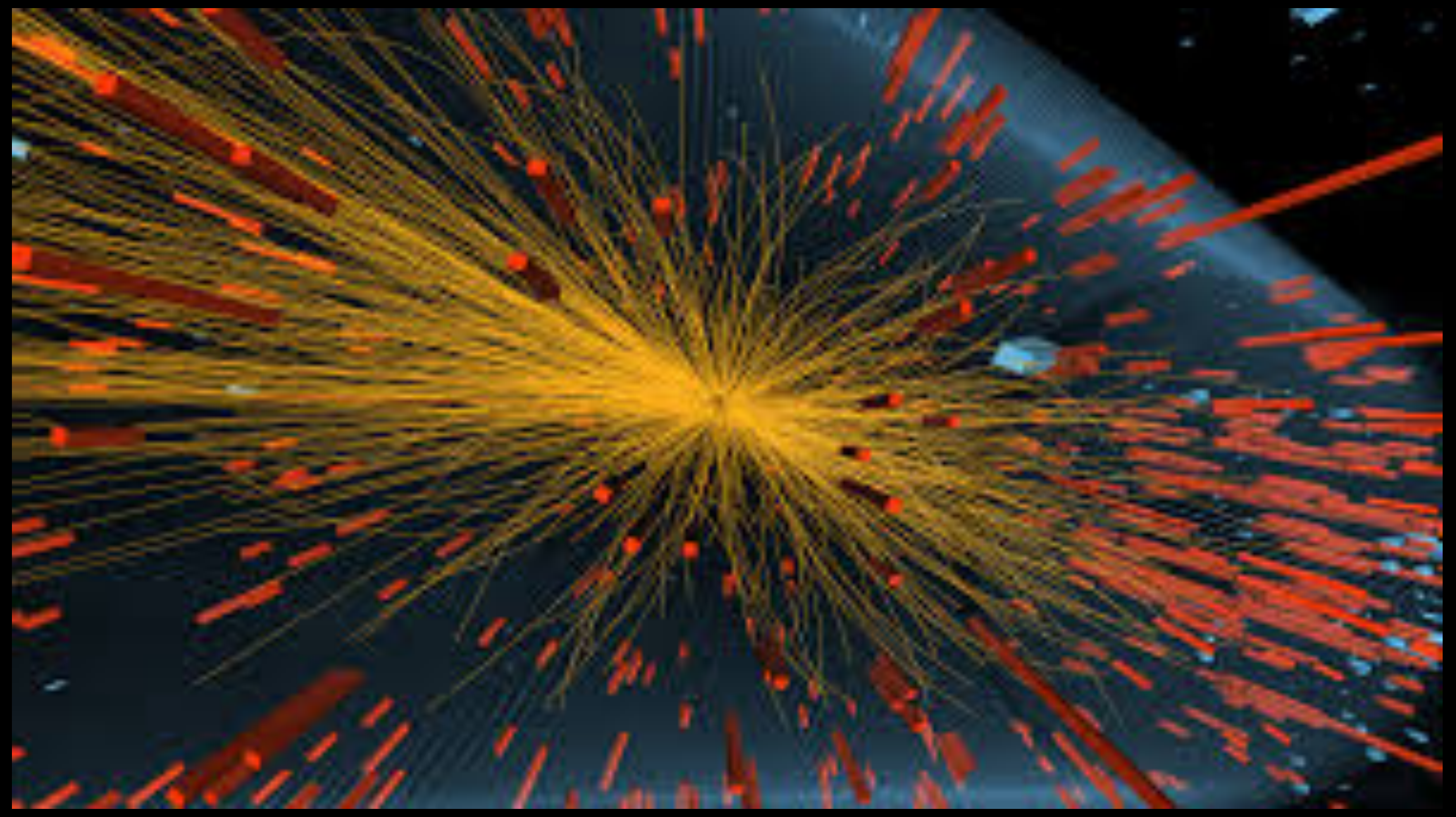




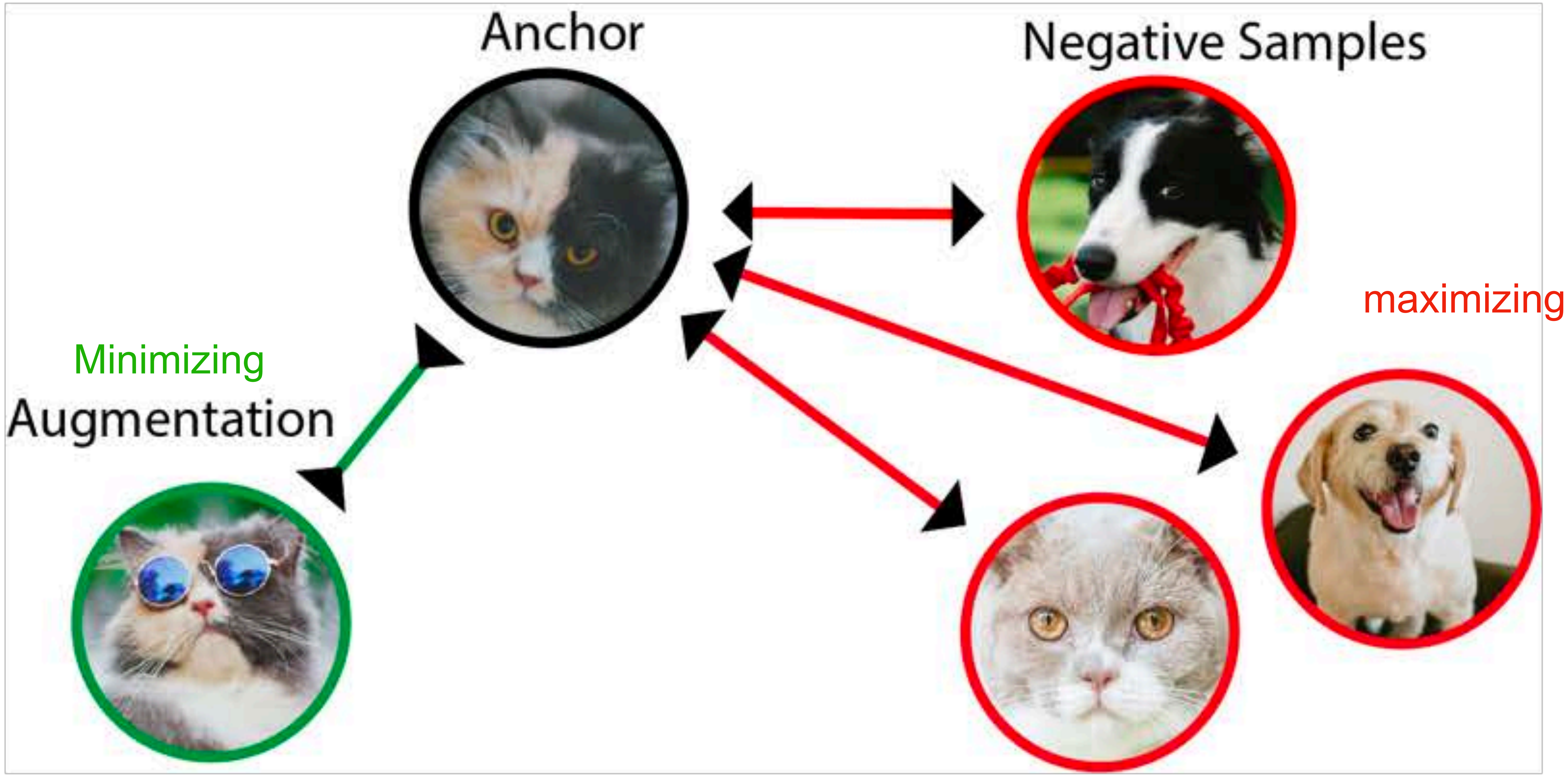
Learning the space

- By looking at data, we can learn a lot
  - Go over input piece by piece
  - Analyze every aspect
  - Compare every feature
- Find distinctive style of the input
  - can be done e.g by looking for a deviation

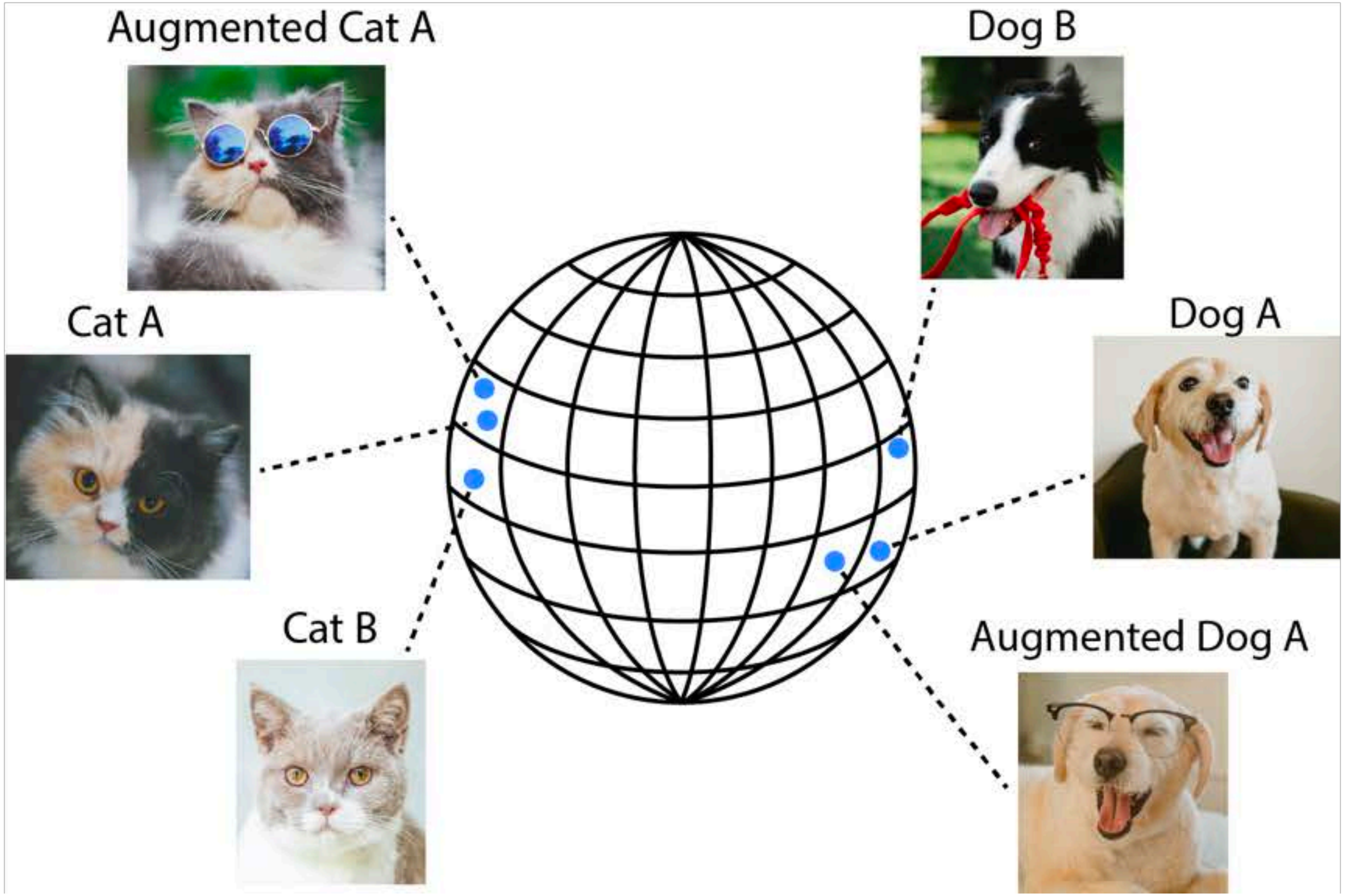




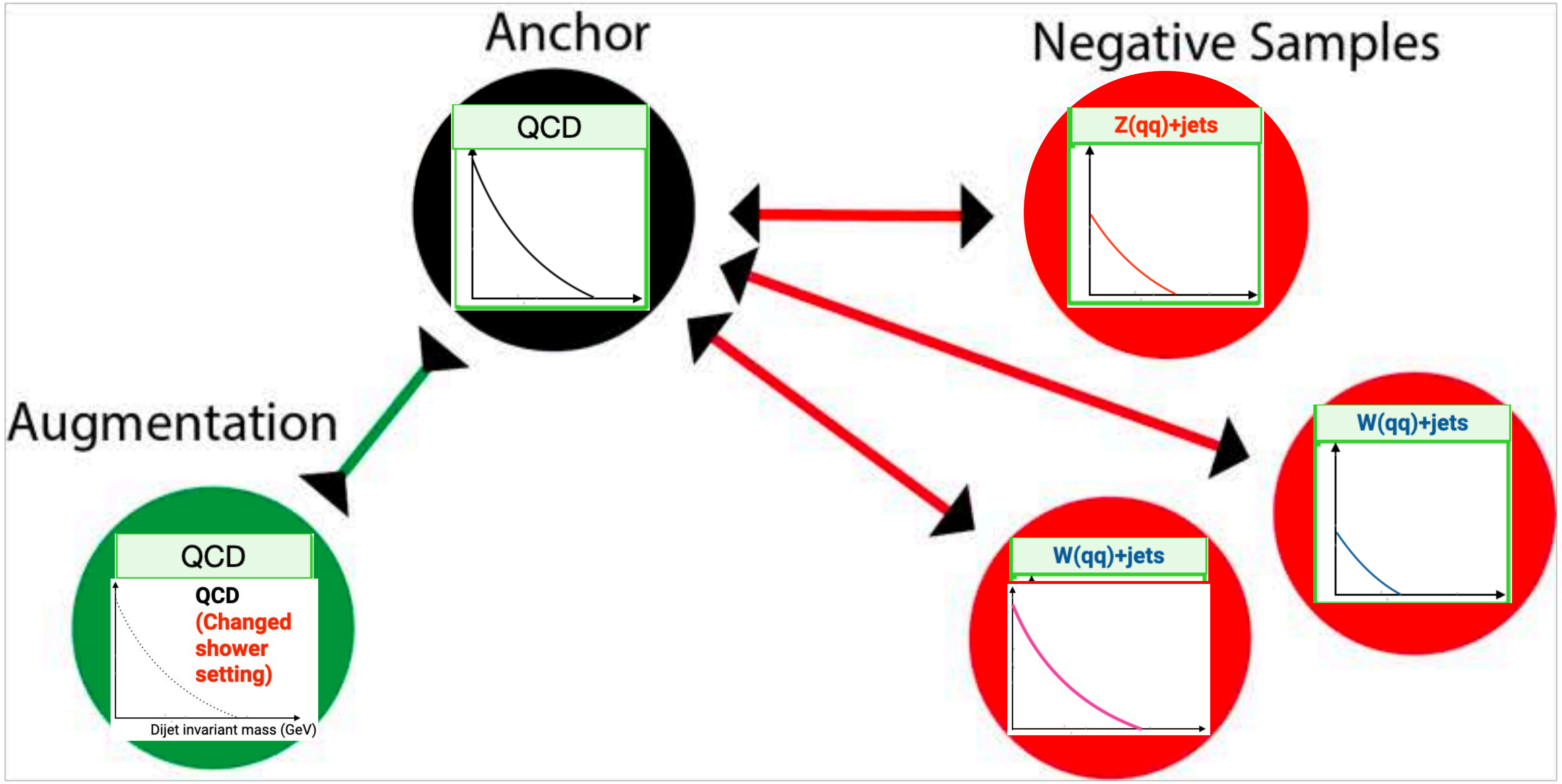
Physically motivated augmentations?



- Minimizing and maximizing distances learns a space

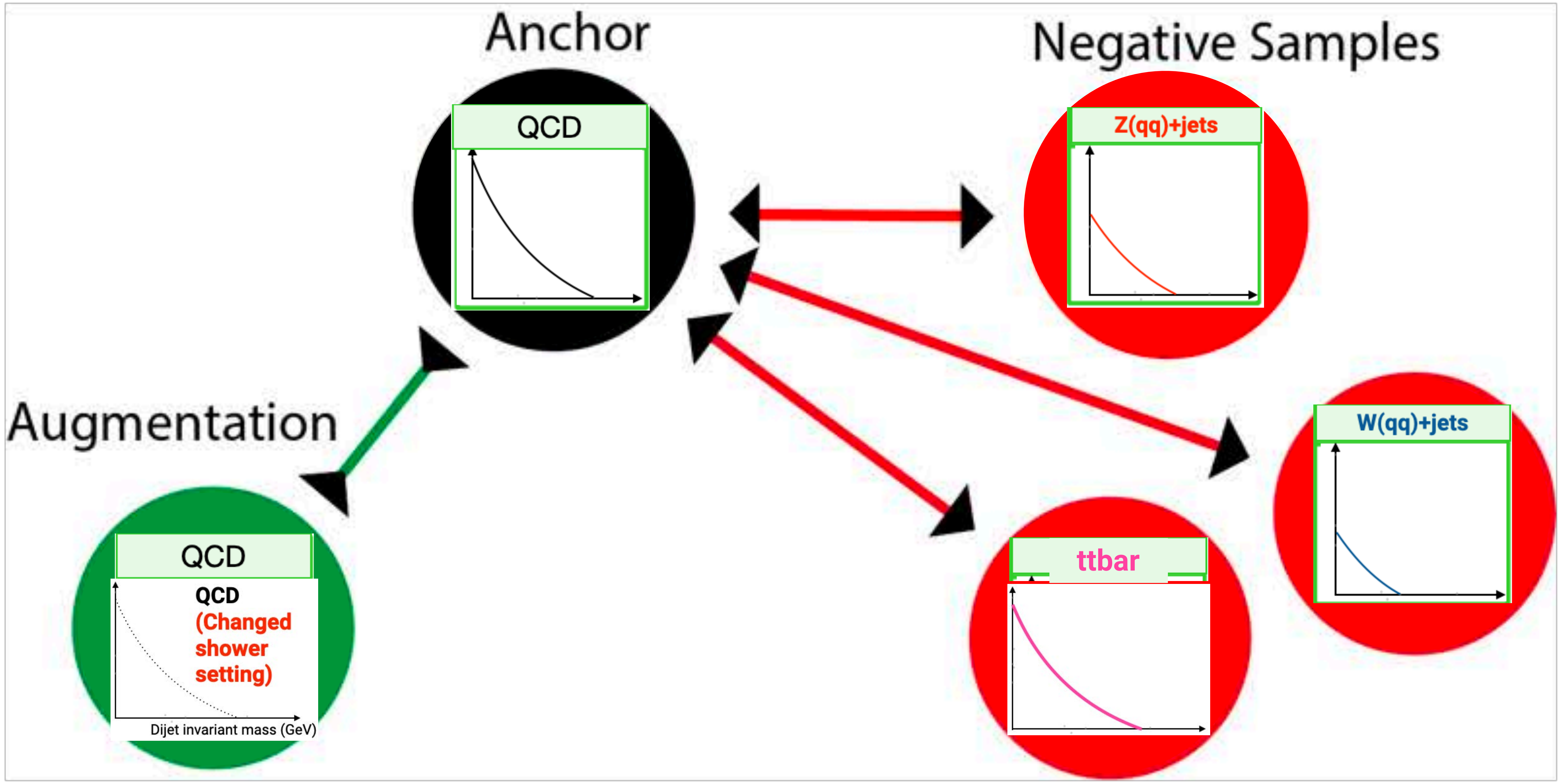

Augmented Cat A
Dog B
Cat A
Dog A
Cat B
Augmented Dog A

Physically motivated augmentations?

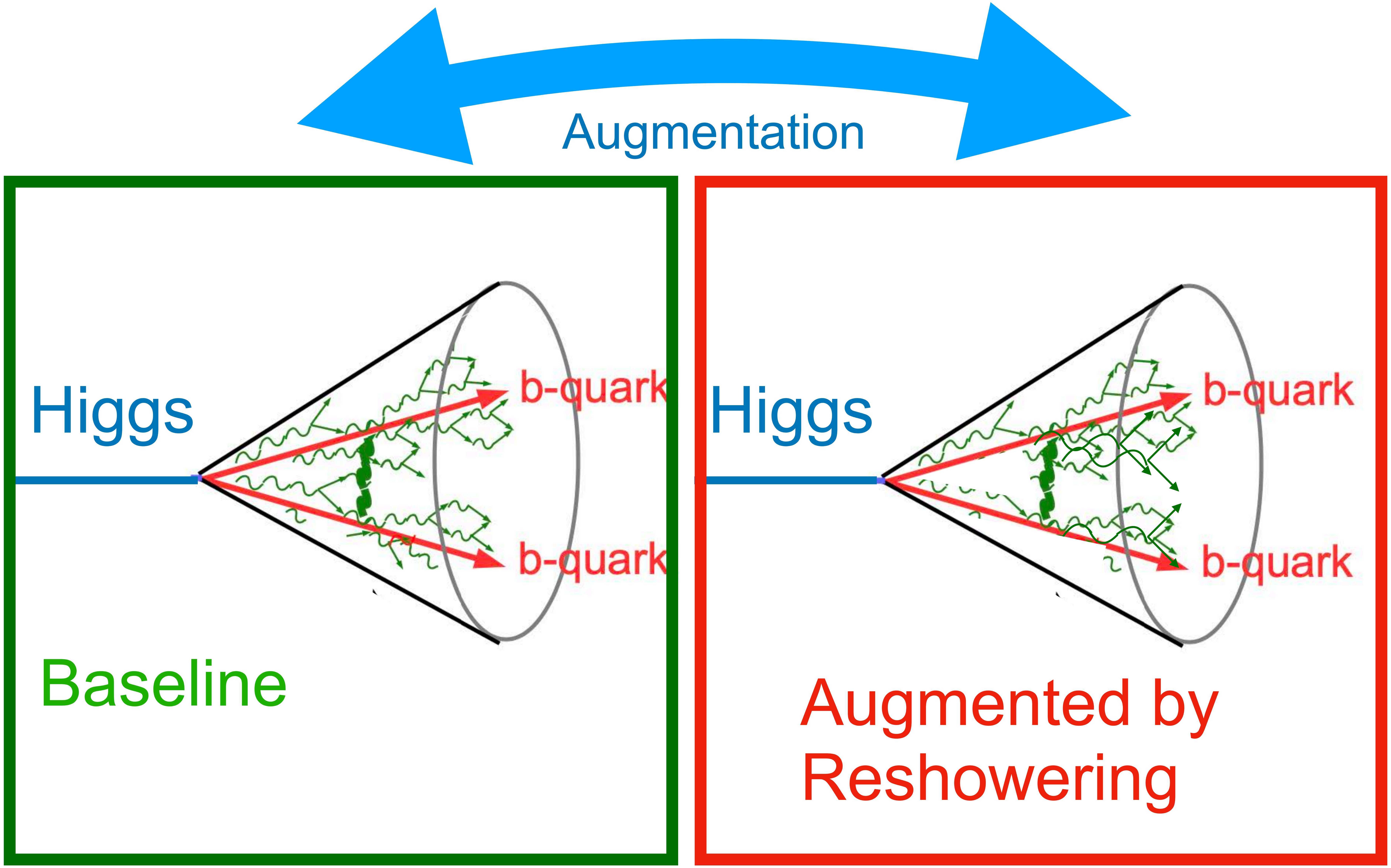


No class labels used in training! How do we augment detector data?

Physically motivated augmentations?

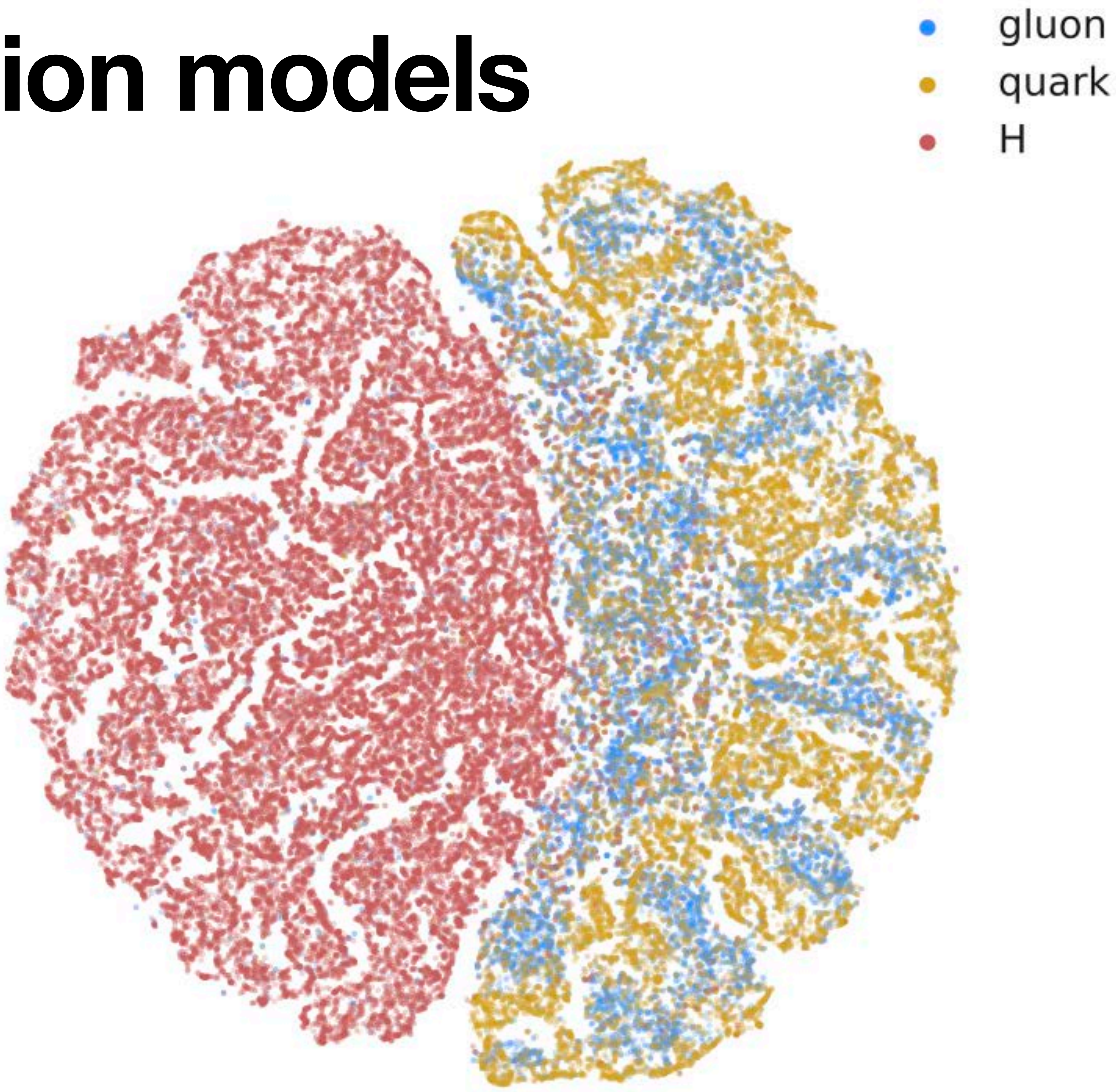


No class labels used in training! How do we augment detector data?

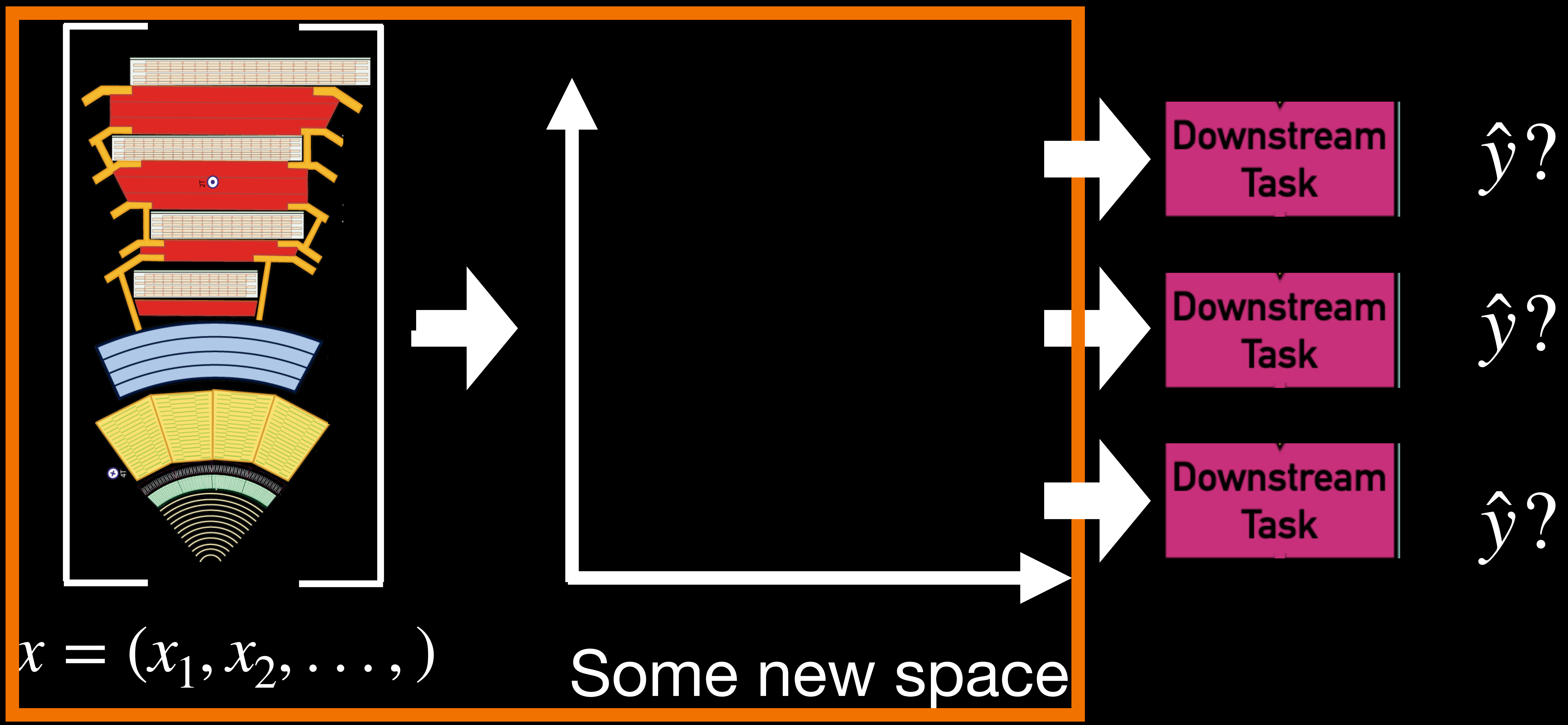


Embedded Space can use any NN to embed

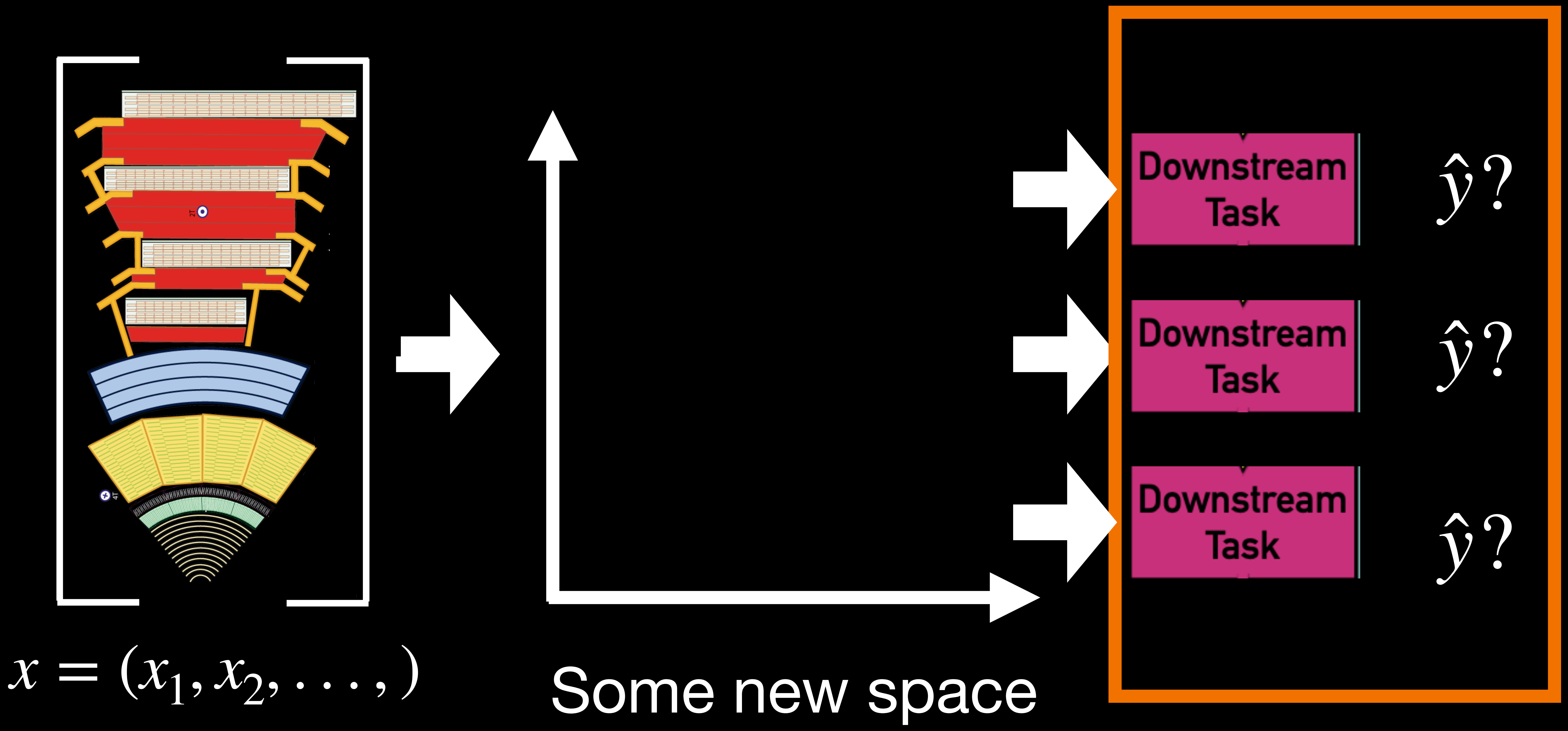
QM foundation models



→ embedding quantum mechanics into AI algorithm

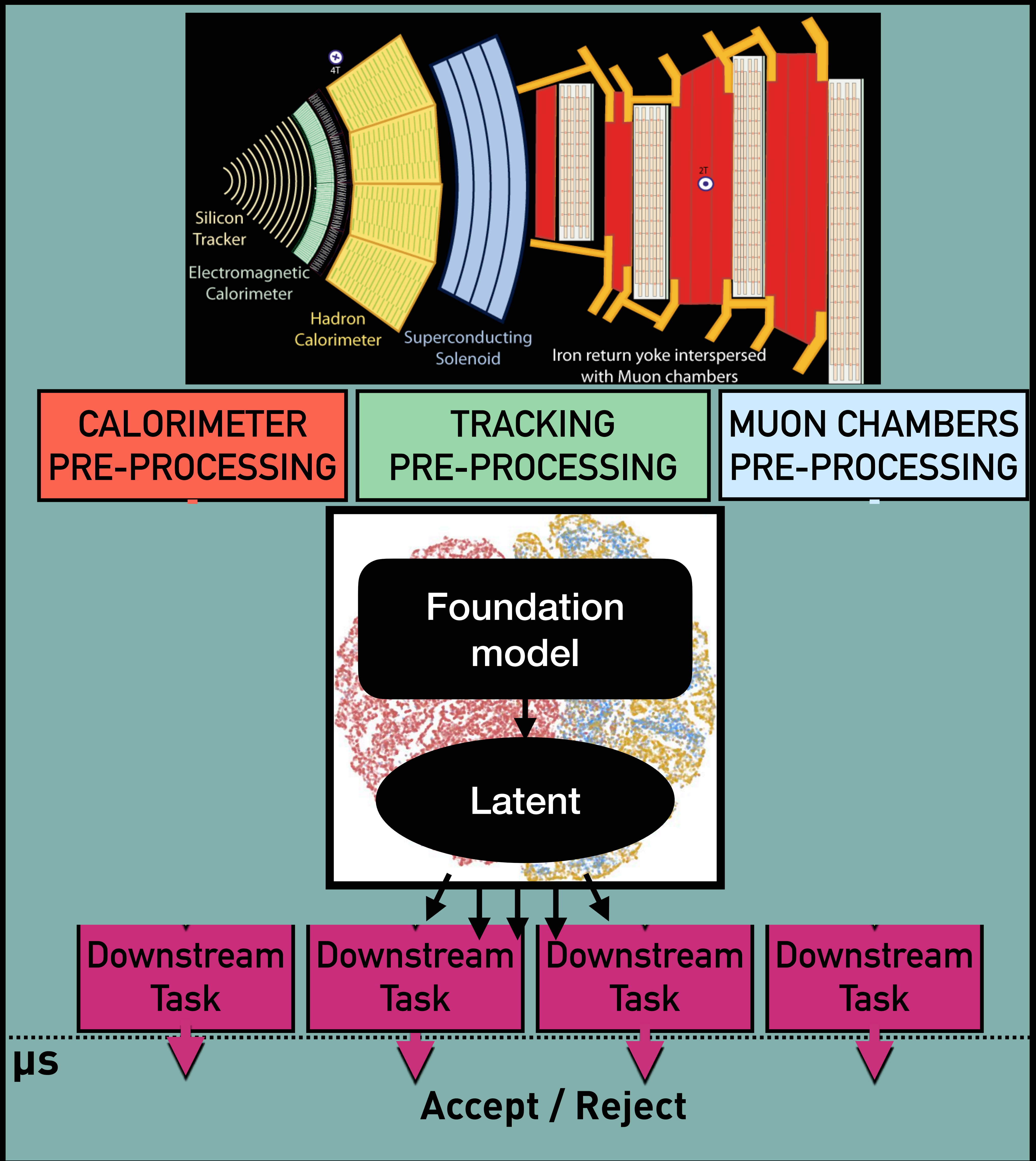


Training 1: Learn neural embedding
(on a lot of data, for a long time)
On simulation? On data?

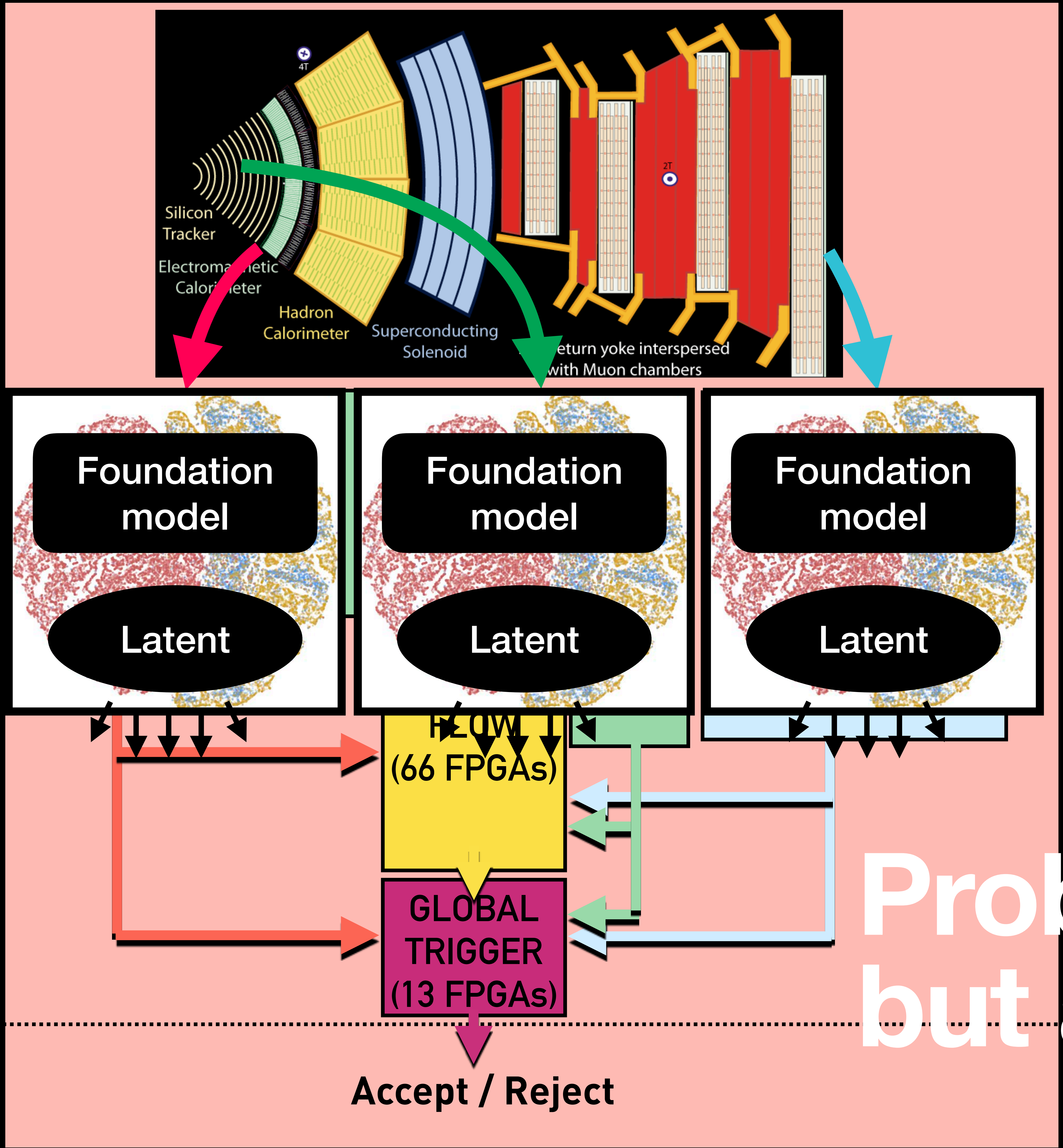


Training 2: Fine tune for specific task (fast, small dataset, simulation)

Foundation model of the Level-1 trigger



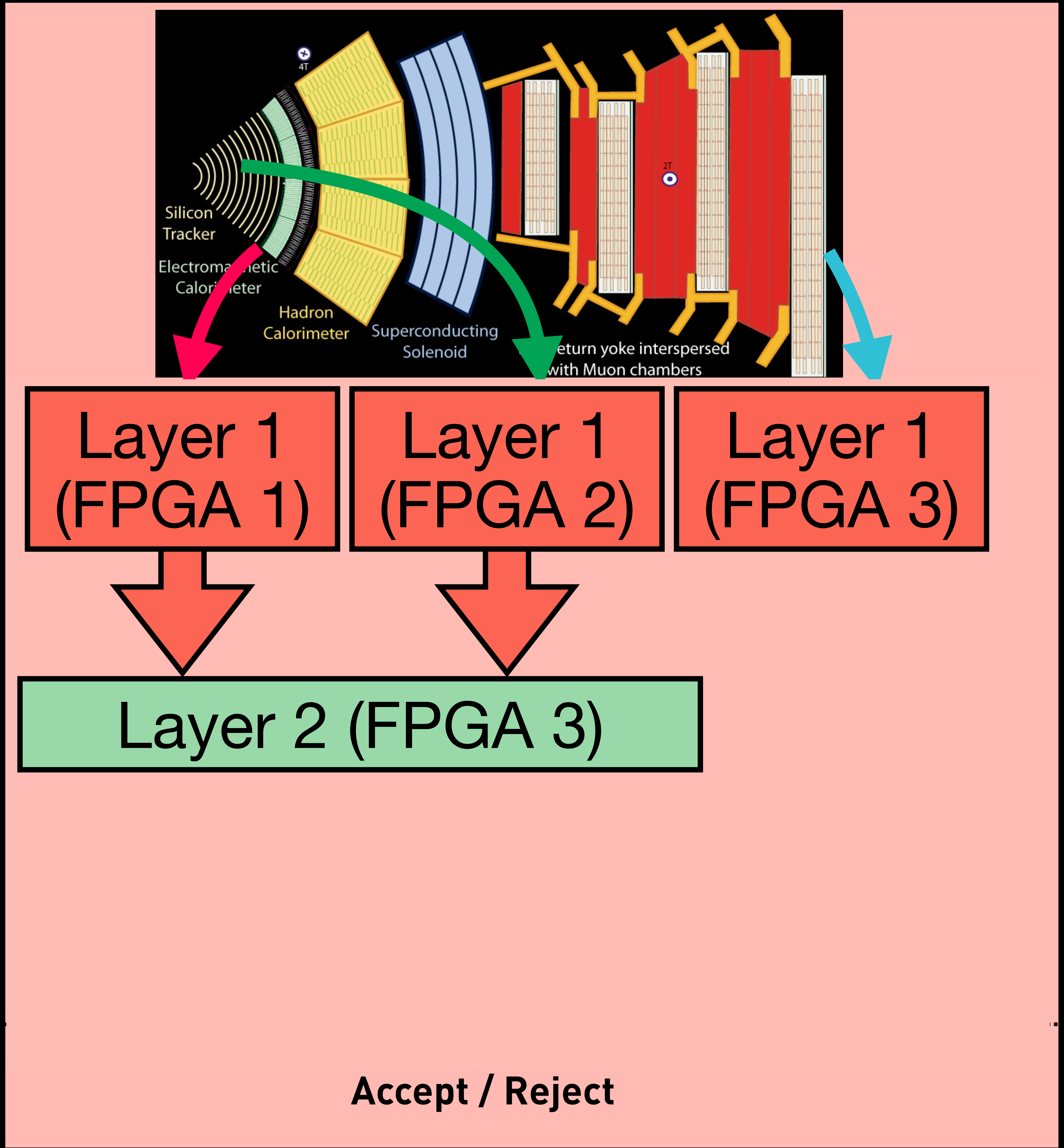
Do I really think this will be possible?



Probably not,
but at some scale

Careful software-hardware co-design

O(1M) parameter model on 1000 FPGAs and do inference in O(1)μs?



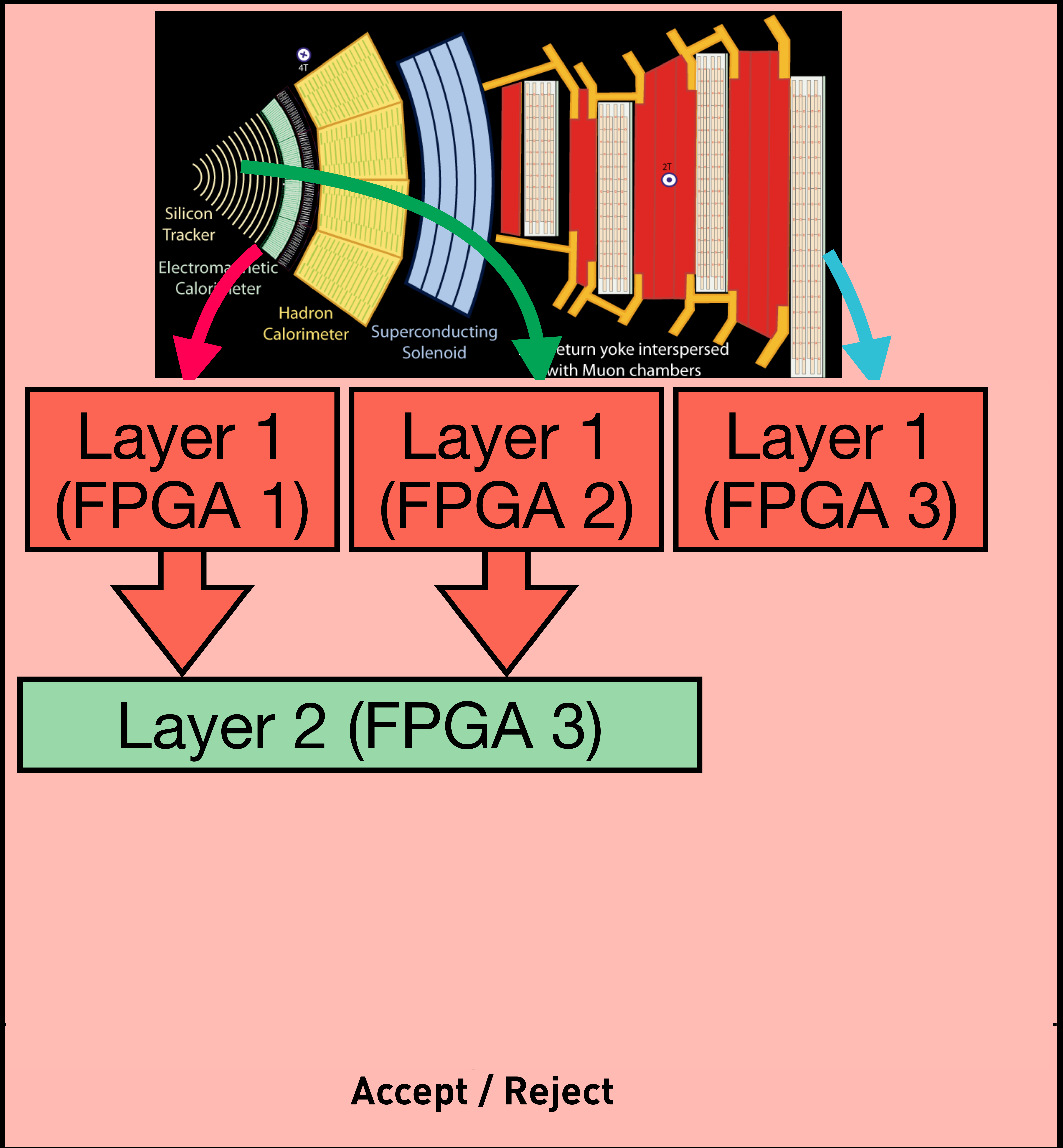
Similar for GPT-4, layers carefully map onto hardware

Algean

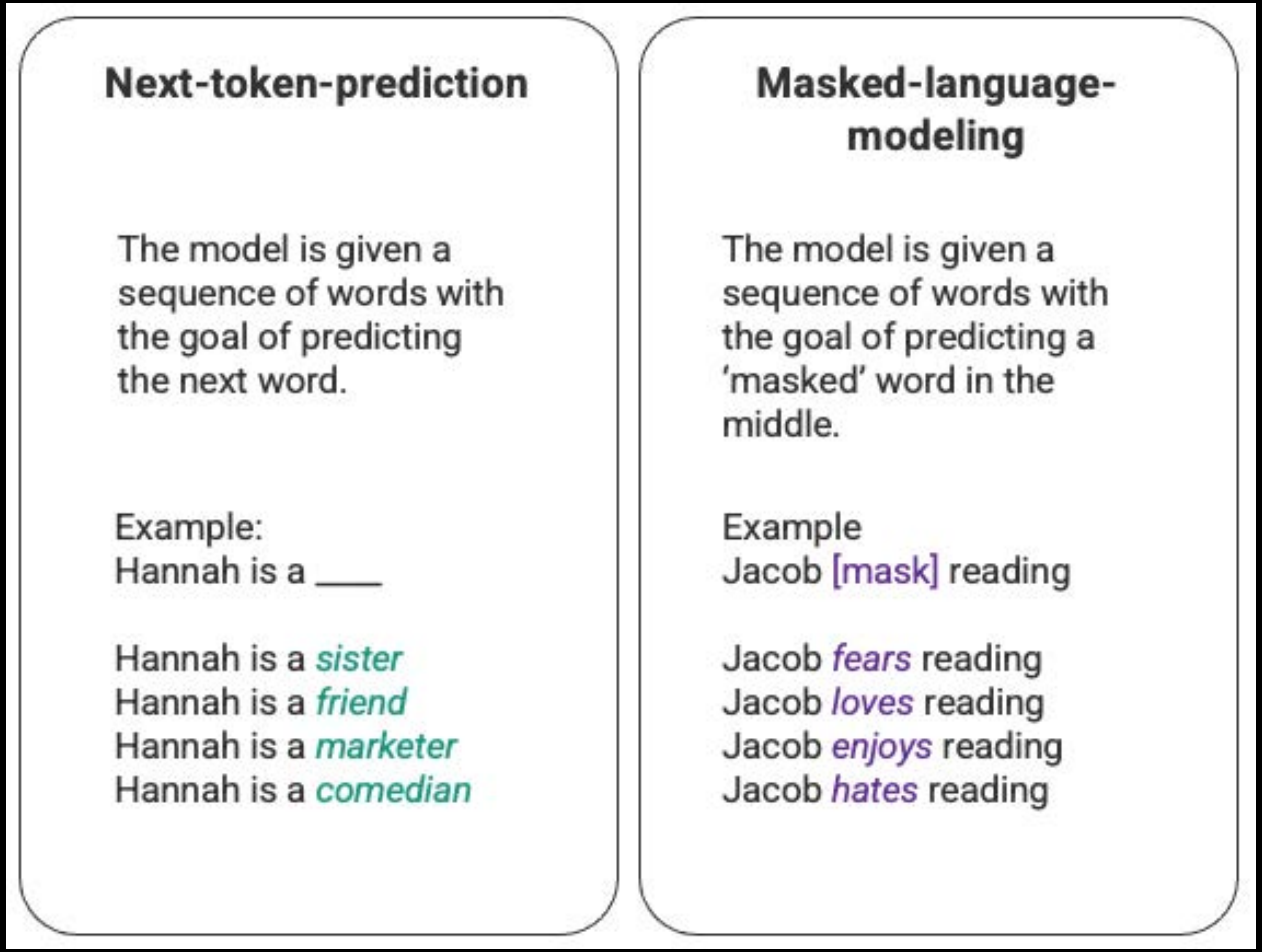
Careful software-hardware co-design

Designed our own protocol to make boards talk to each other fast enough

(25 Gbs to transfer data LHC-synchronously between boards)

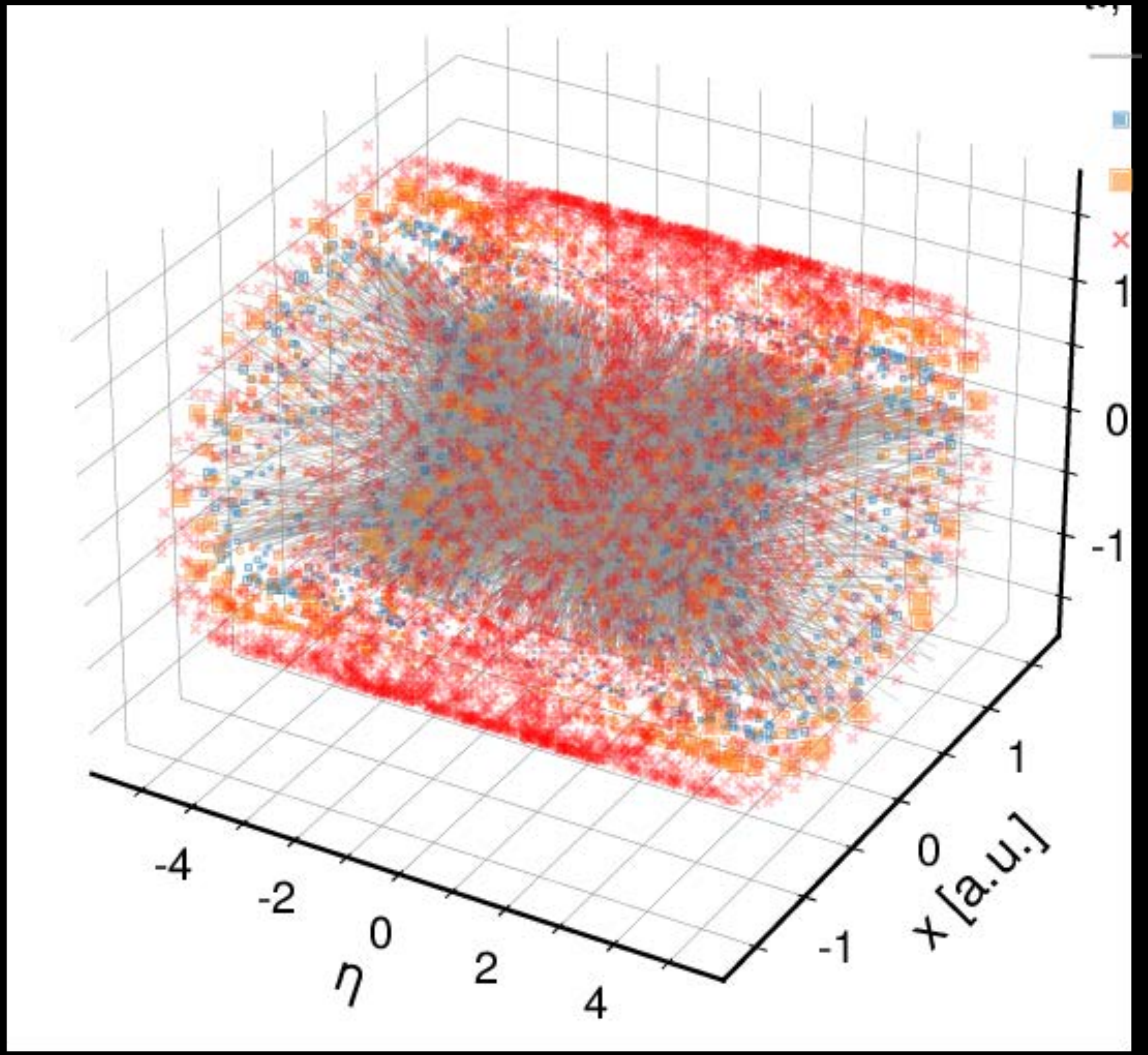


Masked language modelling



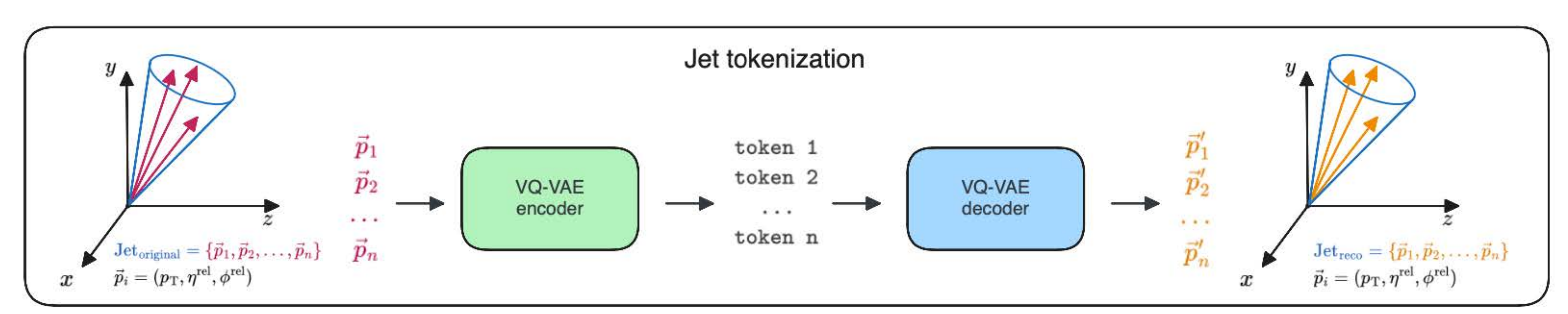
Self-supervised pre-training

Masked particle modelling



Masked calorimeter pre-training?

Tokenisation?



Hardware?


Subscribe to newsletters

Forbes

FORBES > INNOVATION > CLOUD

Groq's Record-Breaking Language Processor Hits 100 Tokens Per Second On A Massive AI Model

Paul Smith-Goodson Contributor
Moor Insights and Strategy Contributor Group

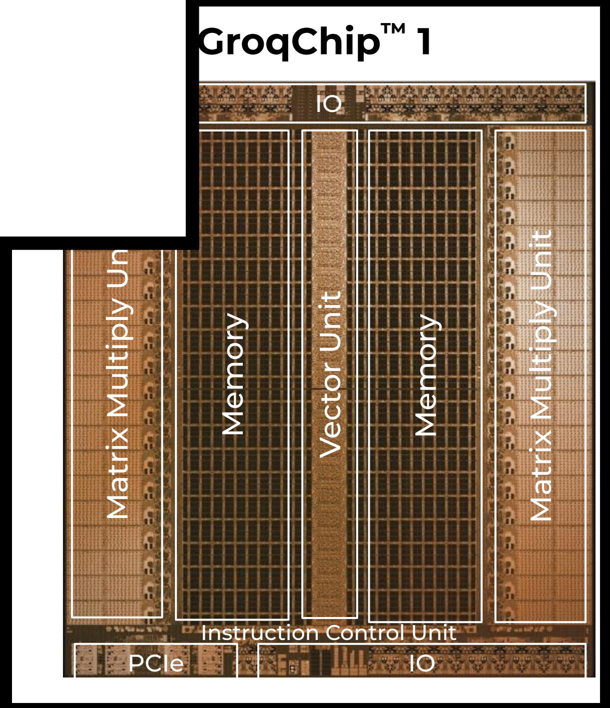
Follow

0

Aug 11, 2023, 03:29pm EDT

0:14 0:00
Ad 1 of 2

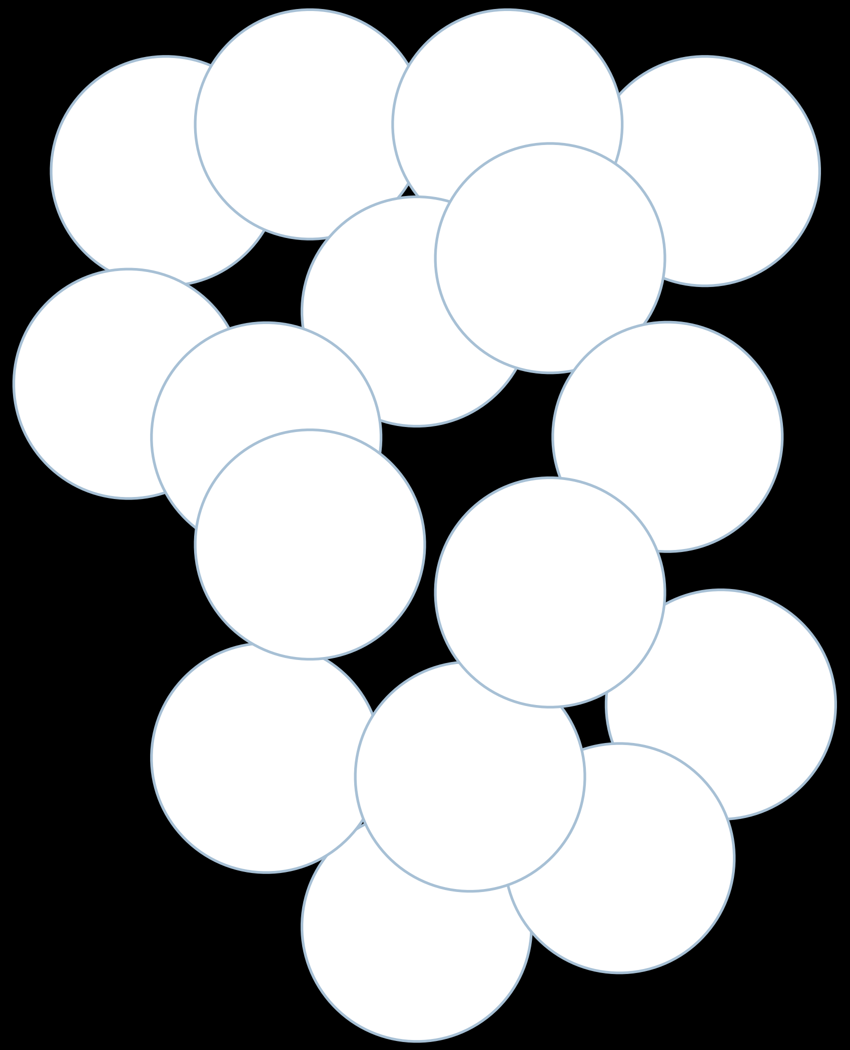
GROQ



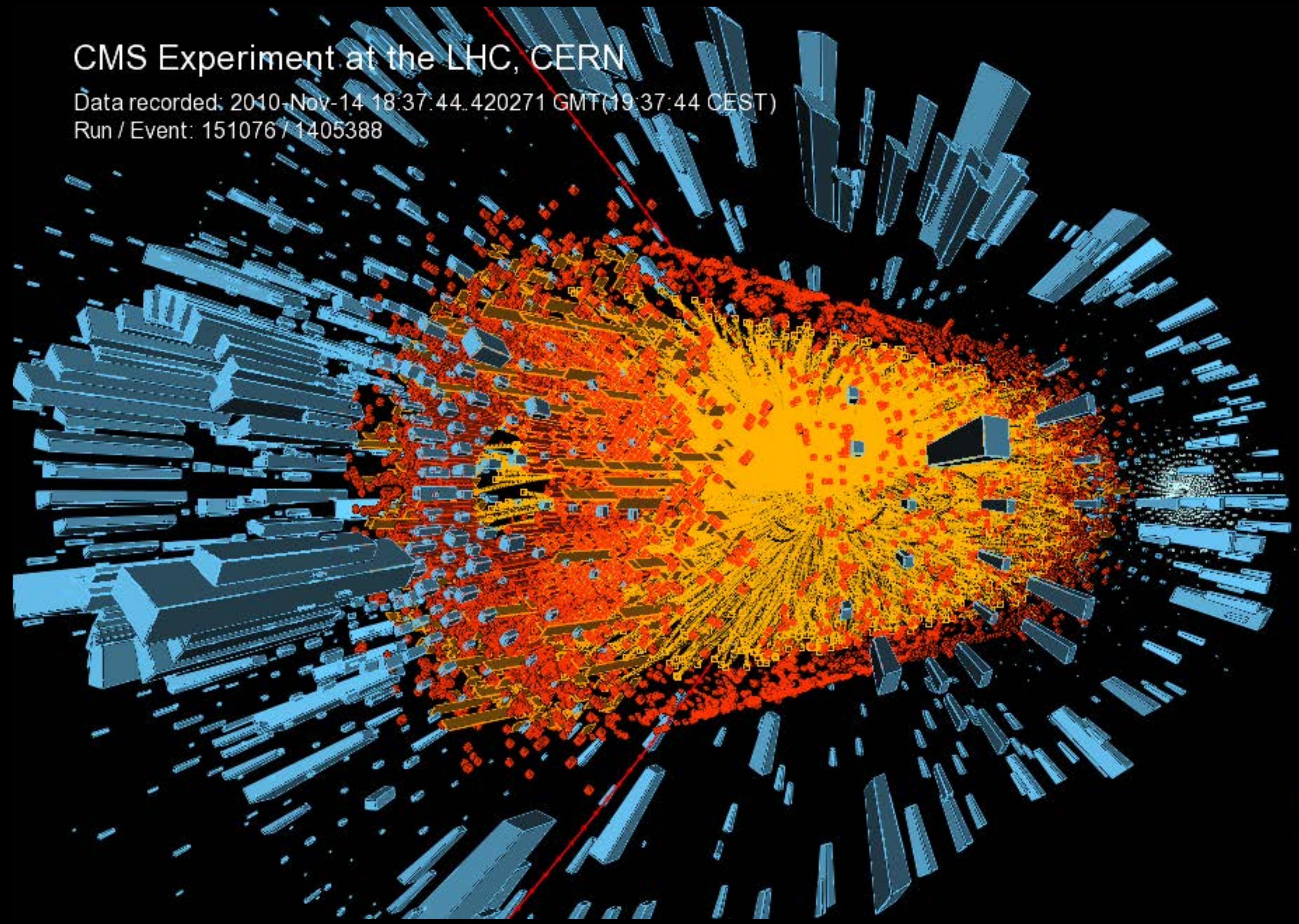
Groq: ultra-low latency dedicated language processor dedicated language processors

- Optimised for sequential data
- First ever 100 tokens/s (usually, ~10)

GPT-4



?


CMS Experiment at the LHC, CERN
Data recorded: 2010-Nov-14 18:37:44.420271 GMT(19:37:44 CEST)
Run / Event: 151076 / 1405388

Backup

Why FPGAs?

Why FPGAs?

- Latency (resource parallelism)

Why FPGAs?

- Throughput (pipeline parallelism)

pipeline

parallelism

Latency (resource parallelism)

Can work on different parts of problem, different data simultaneously

Latency strictly limited by detector frontend buffer

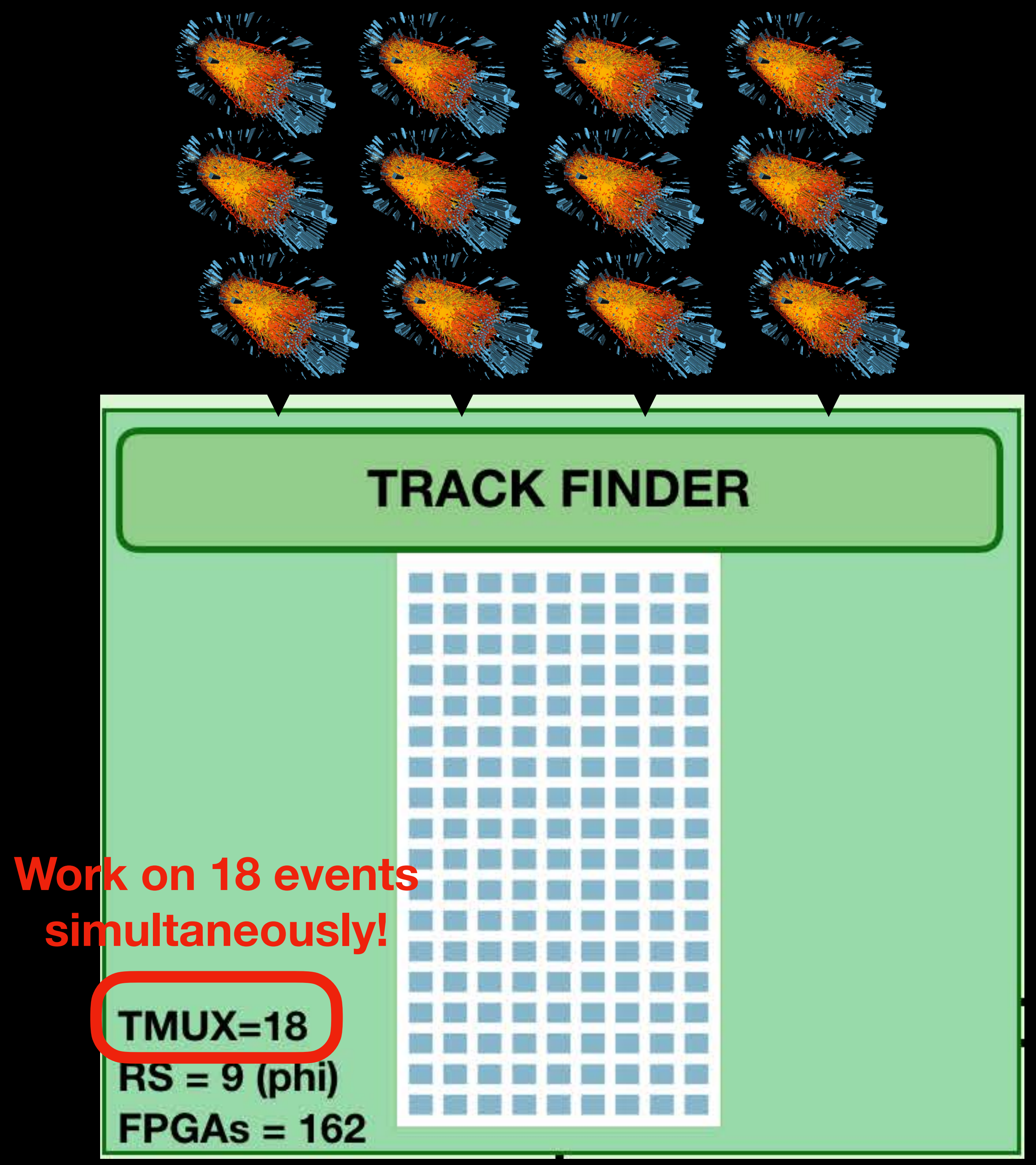
High bandwidth (pipeline parallelism)

Phase 2 L1T processes 5% of total internet traffic

Latency deterministic

CPU/GPU processing randomness,

FPGAs repeatable and predictable latency



$$\mathbf{x}_n = g_n(\mathbf{W}_{n,n-1}\mathbf{x}_{n-1} + \mathbf{b}_n)$$

activation function

precomputed and stored in BRAMs

multiplication

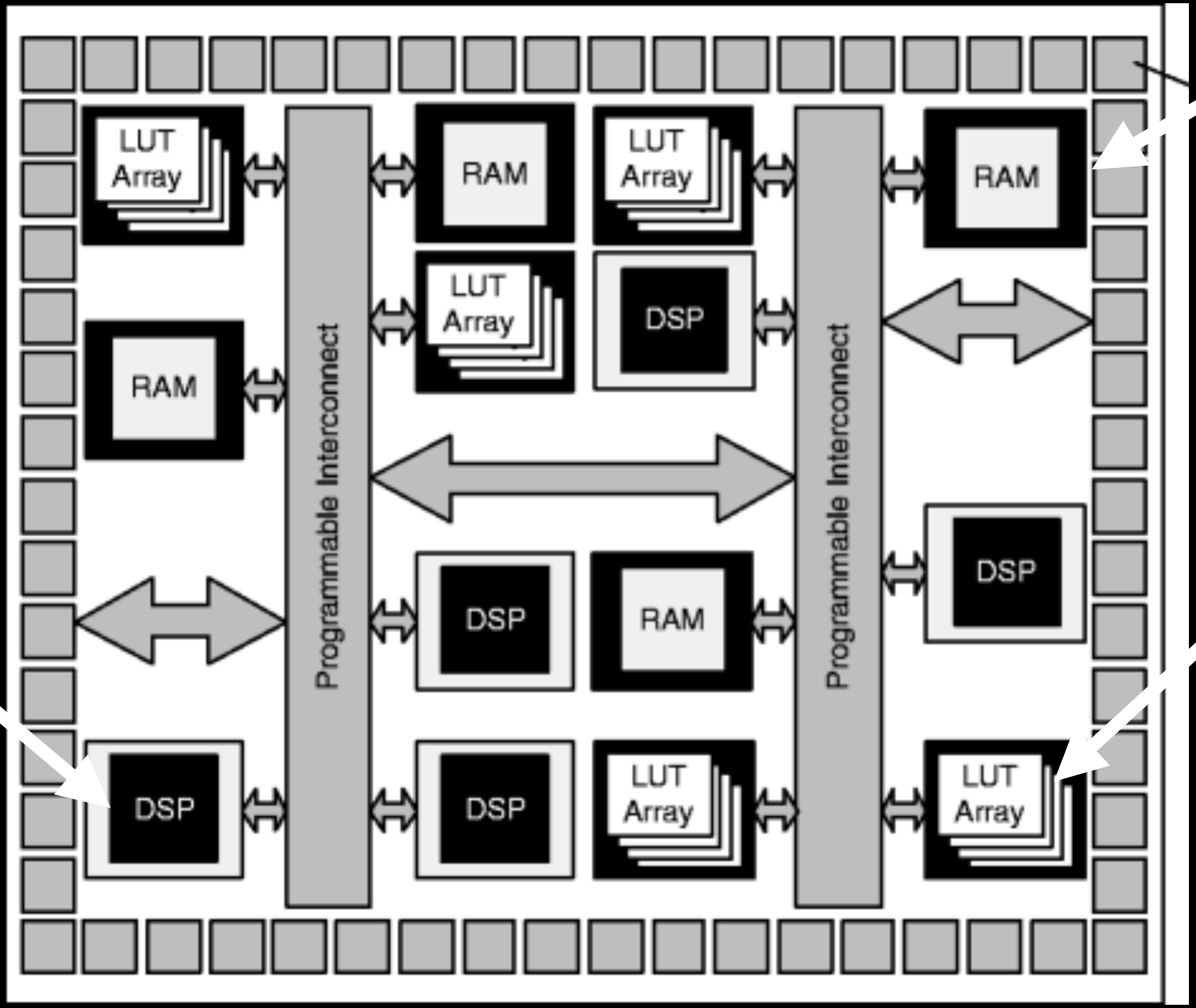
DSPs

addition

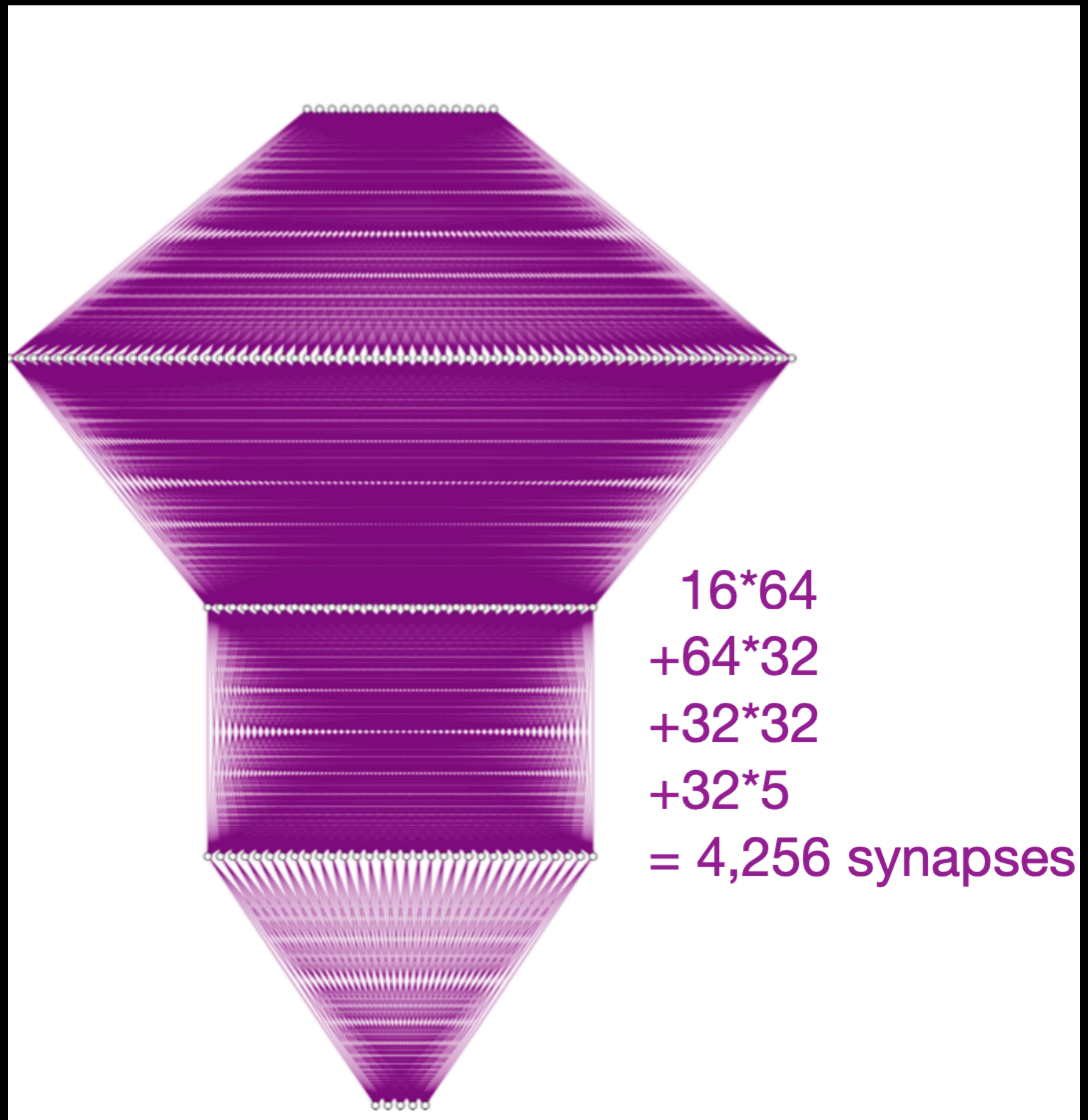
logic cells

Memory (BRAM)
O(2000) units

Digital signal processors (DSPs)
O(5,000) units



Logic cells/lookup tables (LUTs)
O(1) million units



$$\mathbf{x}_n = g_n(\mathbf{W}_{n,n-1}\mathbf{x}_{n-1} + \mathbf{b}_n)$$

activation function

precomputed and stored in BRAMs

multiplication

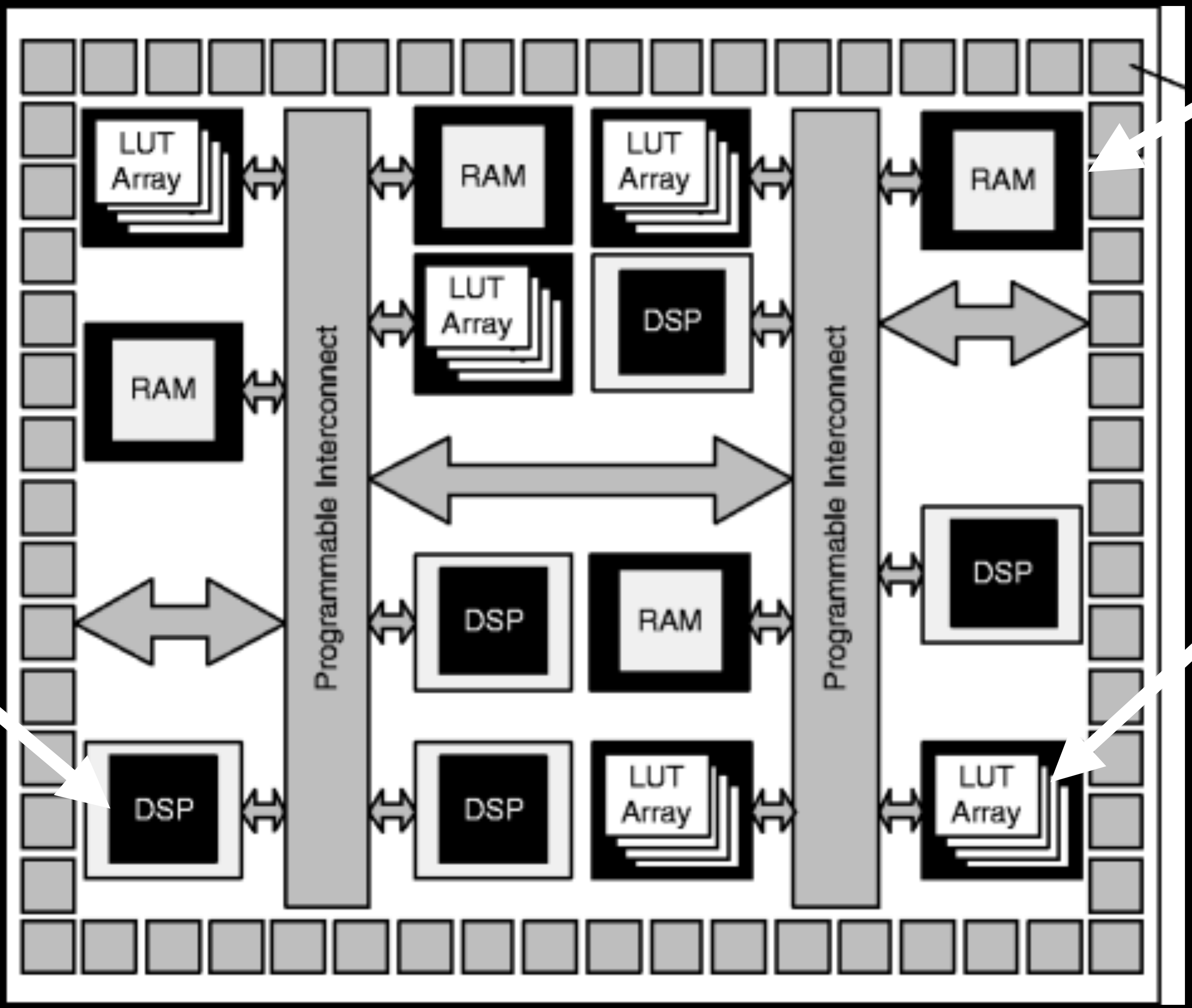
DSPs

addition

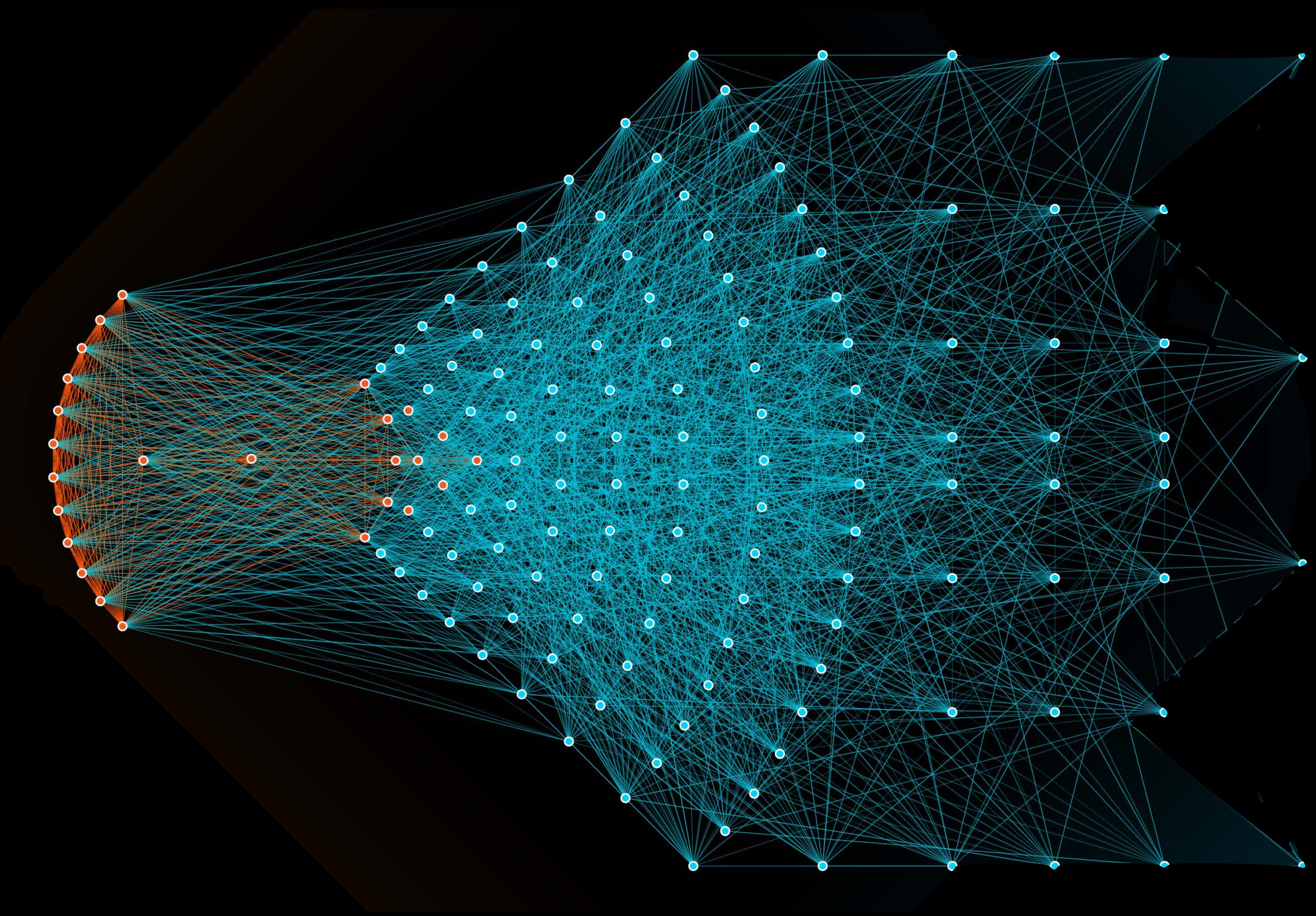
logic cells

Memory (BRAM)
O(2000) units

Digital signal processors (DSPs)
O(5,000) units

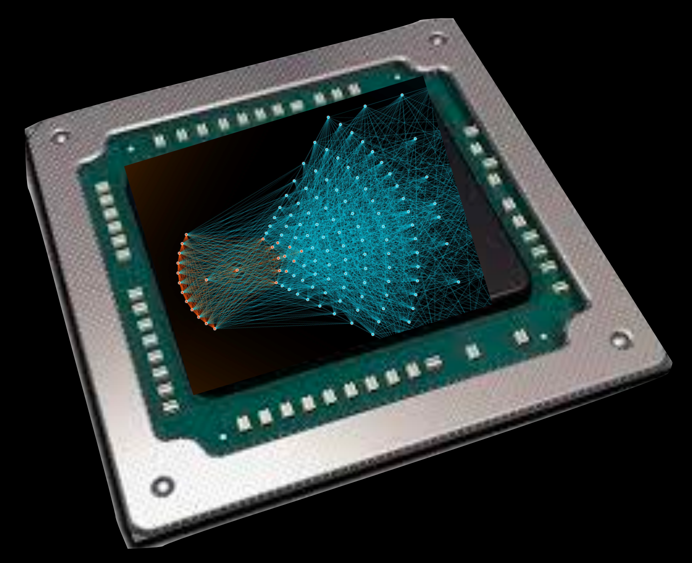


Logic cells/lookup tables (LUTs)
O(1) million units



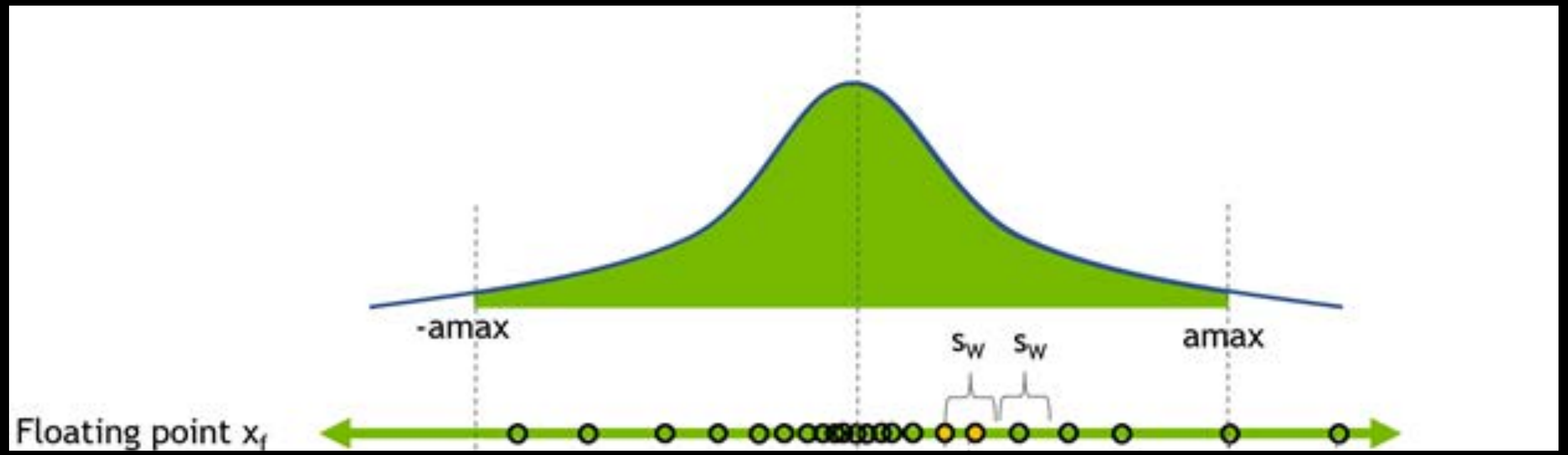
Ideally

- Quantization



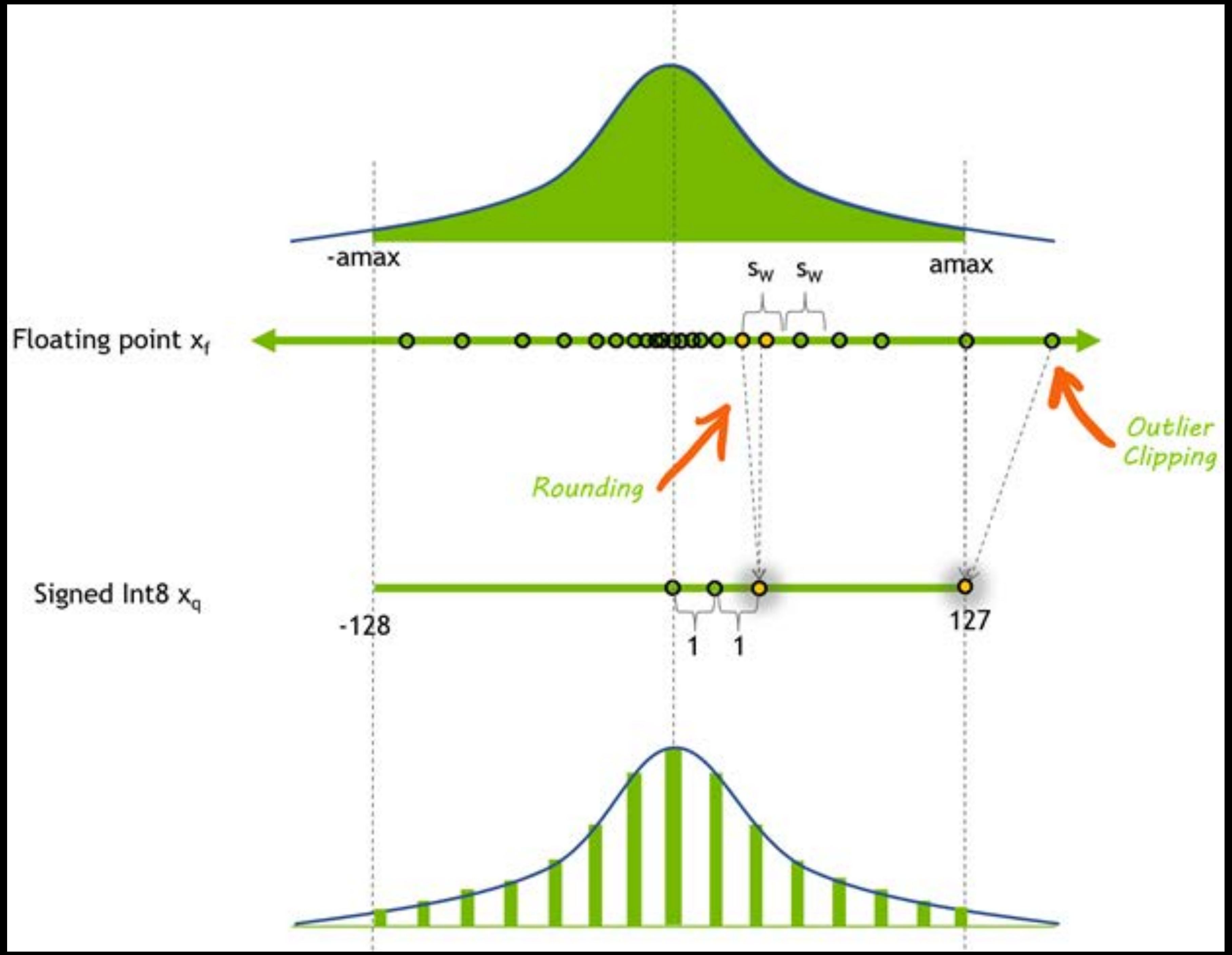
Reality

Quantization



Floating point 32:
4B numbers in [-3.4e38, +3.4e38]

Quantization



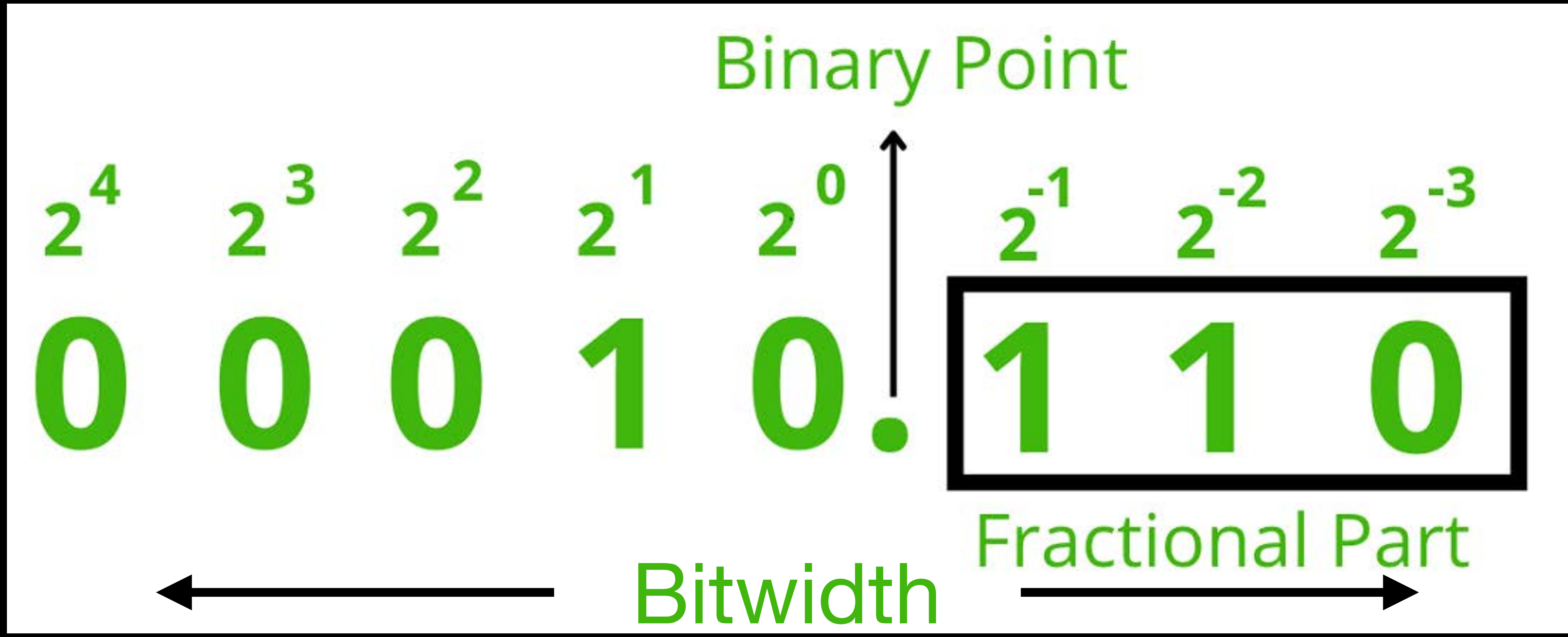
Quantising:

int8 2^8=256 numbers in [-128,127]

$$x_q = Clip(Round(\frac{x_f}{scale}))$$

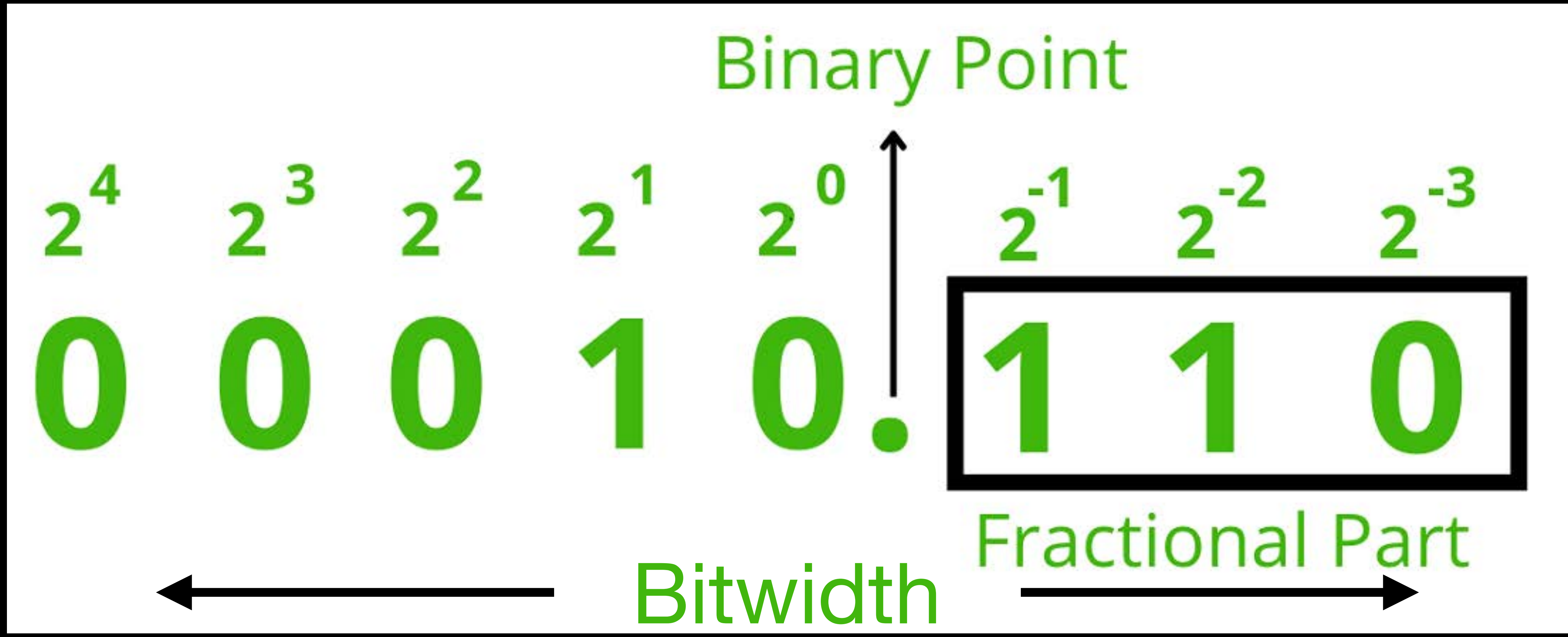
Fixed-point $< W, I >$

a way to express fractions with integers!



$$= 2^4 \cdot 0 + 2^3 \cdot 0 + 2^2 \cdot 0 + 2^1 \cdot 1 + 2^0 \cdot 0 + 2^{-1} \cdot 1 + 2^{-2} \cdot 1 + 2^{-3} \cdot 0 = 2.75$$

Fixed-point $< W, I >$



Trade off: range (integer bits) and precision (fractional bits). E.g $< 8,0 >$:

$$\text{Precision} = \frac{1}{2^{\text{F}}} = \frac{1}{2^8} = 0.00390625$$

$$\text{Range} = [-2^0, -2^0 - 1] = [-1,0]$$

Precision	Approx. **Peak GOPS**	**On-chip weights**
1b	**64 000**	**~64 M**
4b	**16 000**	**~16 M**
8b	**4 000**	**~8 M**
32b	**300**	**~2 M**

200x

30x

Trillions of quantized operations per second

Weights can stay **entirely on-chip**

AMD UltraScale+ MPSoC ZU19EG (conservative estimates)