Invisible Cities

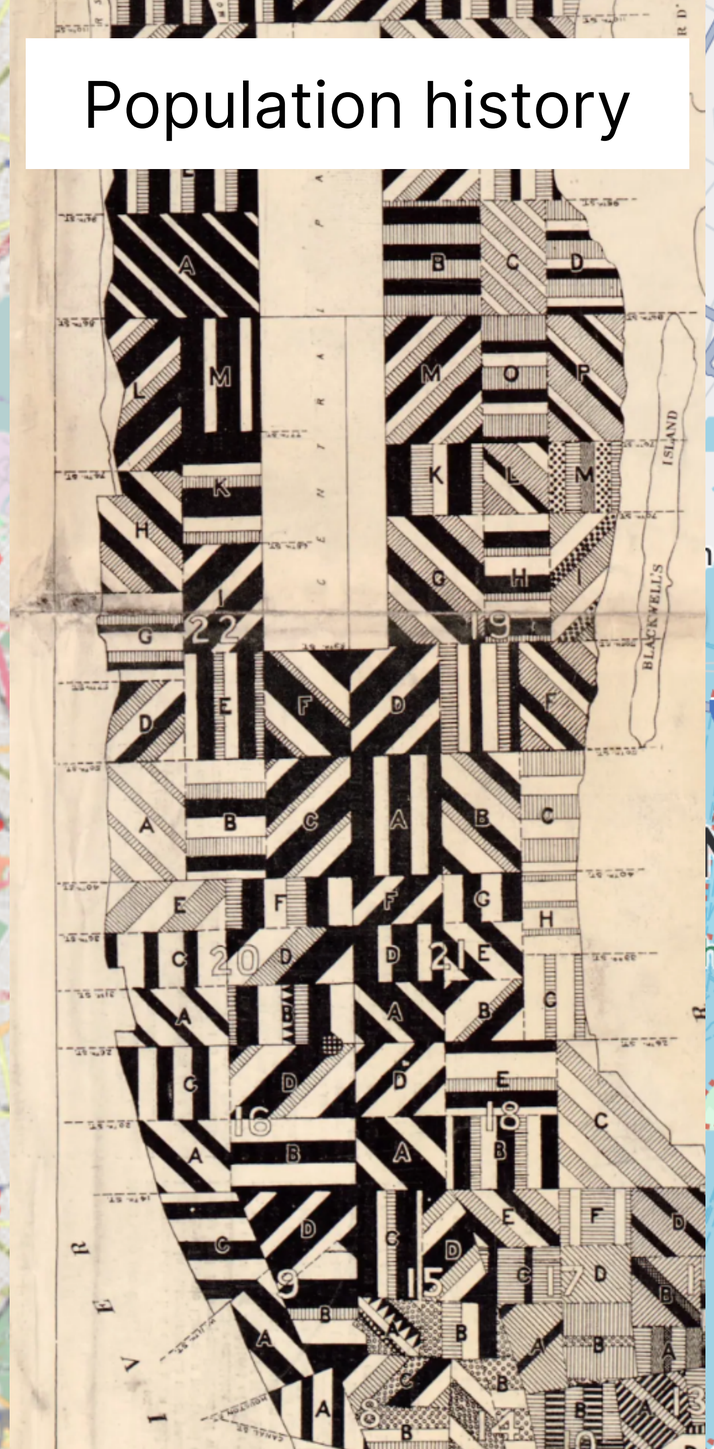
Imagining the next era of AI-enabled fundamental physics research

Mariel Pettee
February 10th, 2026

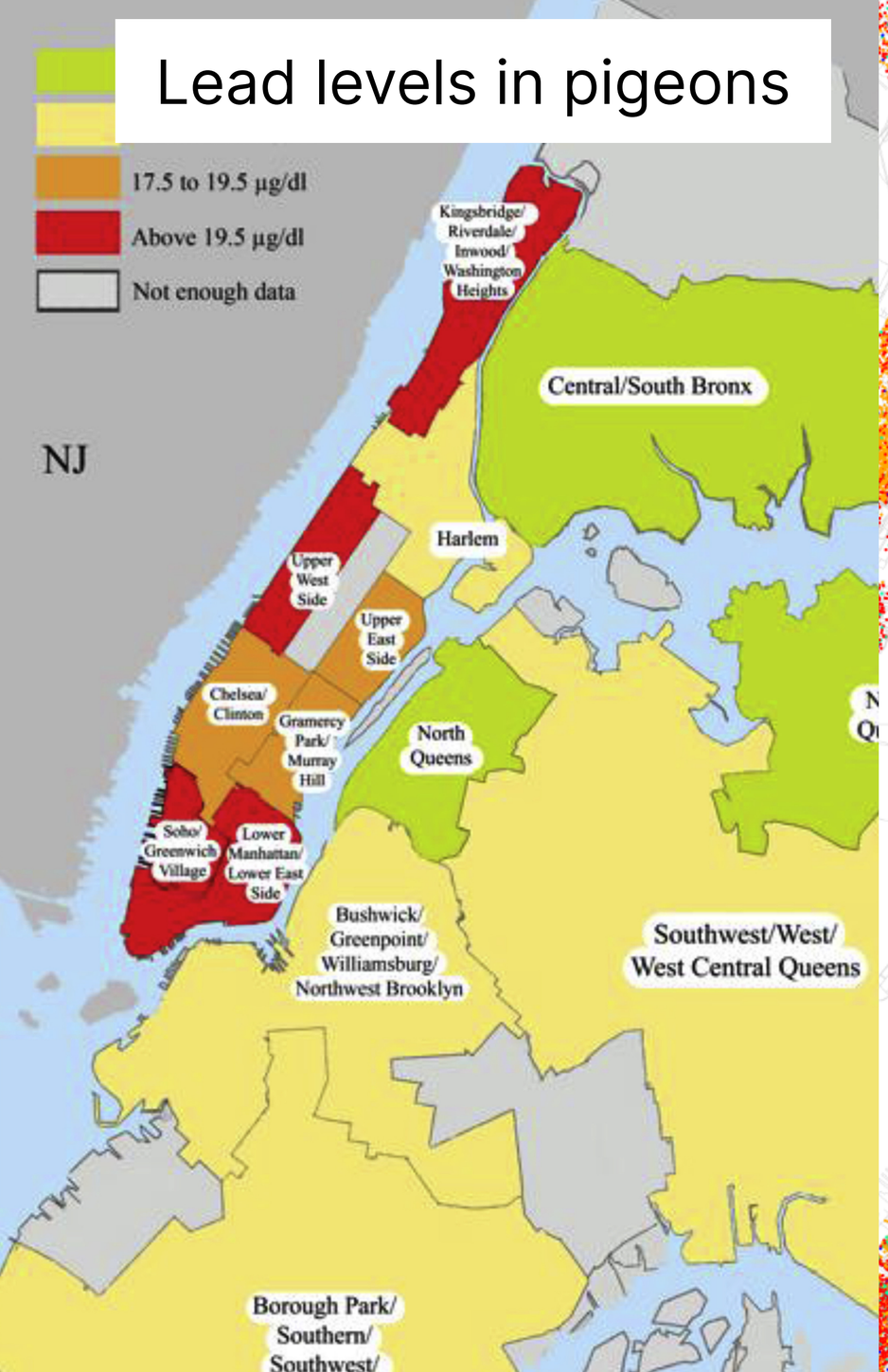

UPPER WEST SIDE
79 St 1
81 St–Museum of Natural History B·C
METROPOLITAN MUSEUM OF ART
UPPER EAST SIDE
77 St 6
72 St B·C
72 St 1·2·3
72 St Q
CENTRAL PARK
68 St Hunter College 6
Lexington Av/63 St F·Q
66 St Lincoln Center 1
Roosevelt Island F
21 St Queens-bridge F
Broadway N·W
LONG ISLAND CITY
36 Av N·W
39 Av N·W
36 St M·R
Steinway St M·R
46 St M·R
Northern Blvd M·R
BROADWAY
rush hour peak direction express
Lexington Av/59 St N·R·W
TRAMWAY
ROOSEVELT ISLAND
Queensboro Plaza N·W·7
59 St Columbus Circle A·B·C·D·1
WEST SIDE
57 St-7 Av N·Q·R·W
57 St F
5 Av/59 St N·R·W
59 St 4·5·6
QUEENSBORO BRIDGE
Court Sq-23 St E·M
Lexington Av/53 St E·M
Queens Plaza E·M·R
33 St-Rawson St 7
40 St Lowery St 7
46 St Bliss St 7
52 St 7
50 St southbound C·E
50 St 1
49 St N·R·W northbound
5 Av/53 St E·M
51 St 6
Court Sq 7 G·7
42 St/Port Authority Bus Terminal A·C·E
47–50 Sts Rockefeller Ctr B·D·F·M
UNITED NATIONS
21 St G
Hunterspoint Av
Hunters Point Av 7·LIRR
LONG ISLAND EXPWY
GREENPOINT AV
JAVITS CENTER
Times Sq-42 St N·Q·R·S·W·1·2·3·7 except S
42 St Bryant Pk B·D·F·M
5 Av 7
Grand Central 42 St except S·4·5·6·7·Metro-North
Vernon Blvd Jackson Av 7
QUEENS MIDTOWN TUNNEL
34 St Hudson Yards 7
34 St Penn Station A·C·E·LIRR
34 St Penn Station 1·2·3·LIRR
34 St Herald Sq B·D·F·M N·Q·R·W
33 St 6
28 St R·W
28 St 6
28 St 1
Long Island City
Greenpoint Av G
GREENPOINT
MANHATTAN AV
QUEENS BROOKLYN
CHELSEA
23 St C·E
23 St 1
23 St F·M
23 St R·W
23 St 6
14 St–Union Sq L·N·Q·R·W·4·5·6
L·N·Q·R·W
Nassau Av G
NASSAU AV
Bedford Av L
18 St 1
8 Av L
6 Av L
3 Av L
1 Av L
N 7 ST
Lorimer St L
Graham Av L
14 St A·C·E
14 St 1·2·3
14 St F·M
Astor Pl 6
8 St-NYU R·W
EAST VILLAGE
E 8 ST
Metropolitan Av G
WILLIAMSBURG
WASHINGTON SQUARE PARK
GREENWICH PATH AV
BLEECKER ST
WEST ST
BOWERY
LAFAYETTE
East River
Amtrak

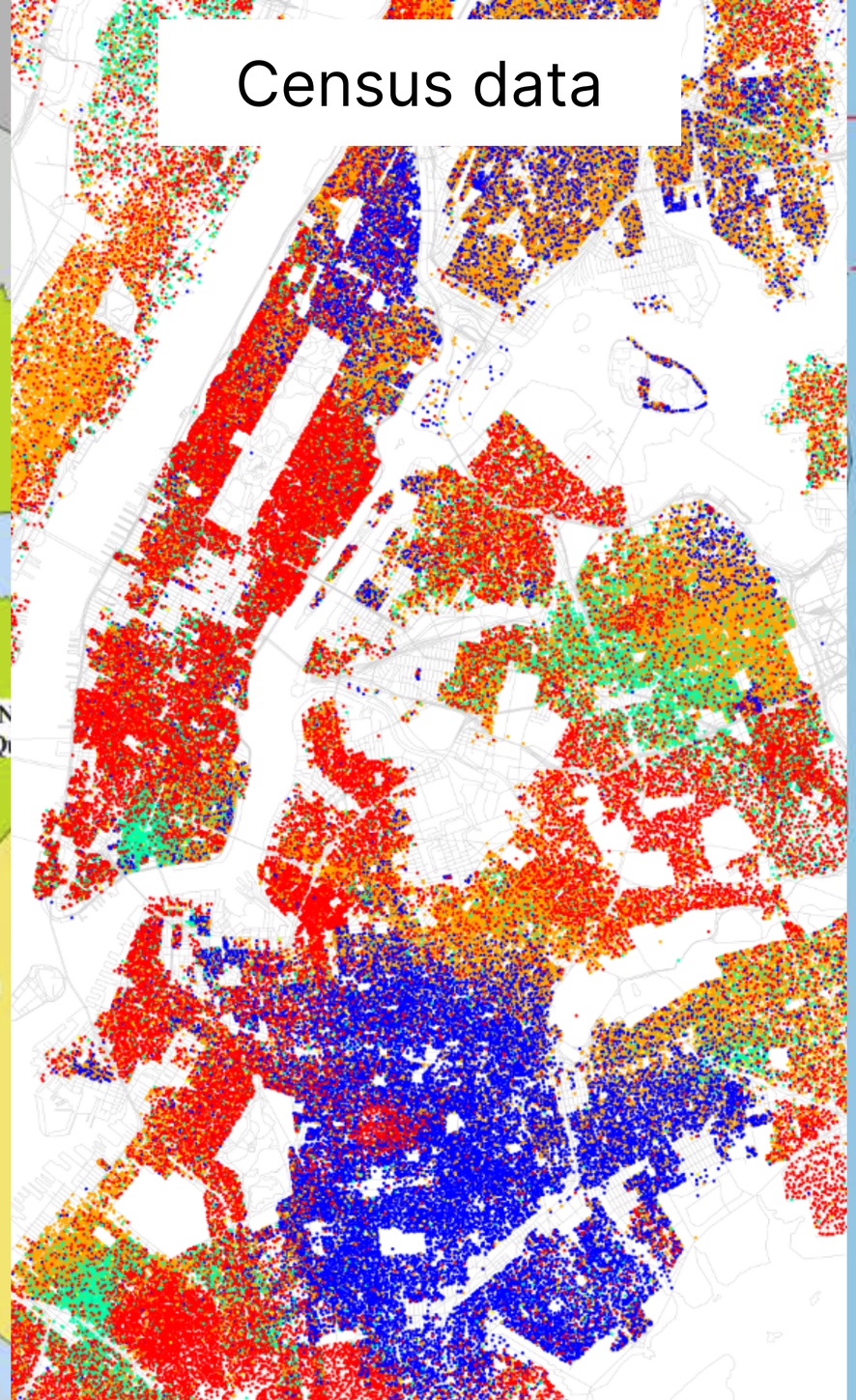
Highest-rated pizza

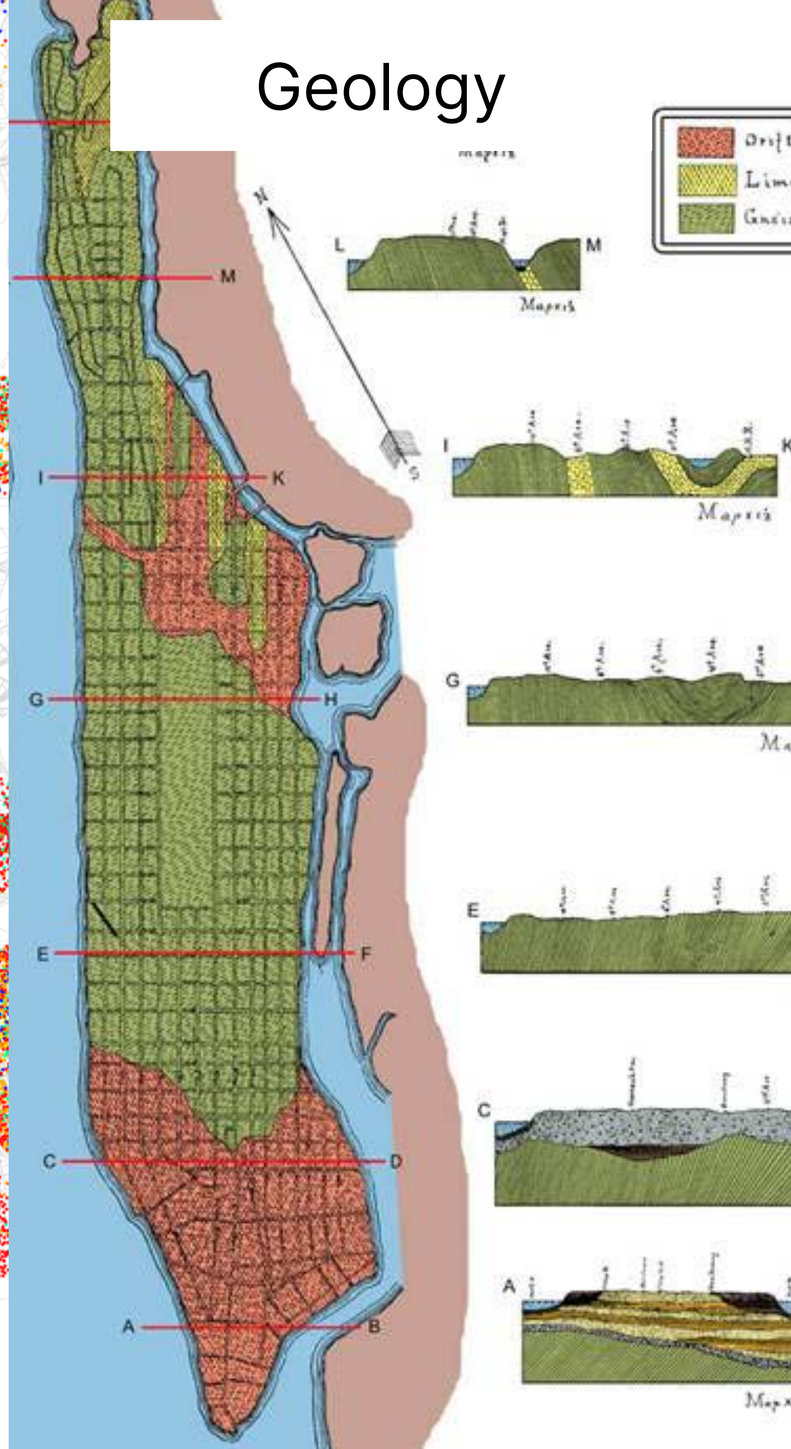
Population history



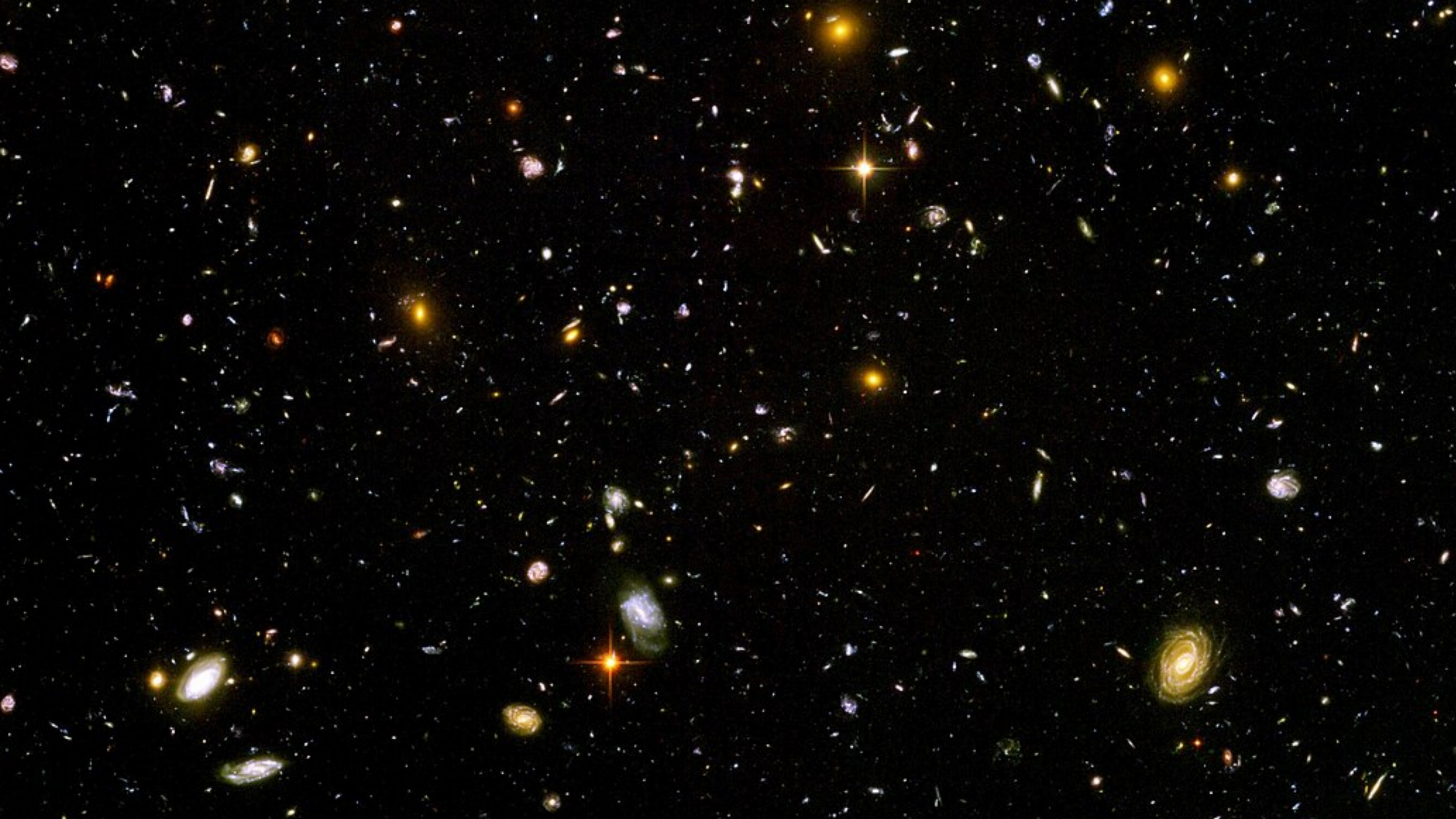
Google Maps


Lead levels in pigeons
17.5 to 19.5 µg/dl
Above 19.5 µg/dl
Not enough data
NJ
Kingsbridge/ Riverdale/ Inwood/ Washington Heights
Central/South Bronx
Harlem
Upper West Side
Upper East Side
Chelsea/ Clinton
Gramercy Park/ Murray Hill
North Queens
Soho/ Greenwich Village
Lower Manhattan/ Lower East Side
Bushwick/ Greenpoint/ Williamsburg/ Northwest Brooklyn
Southwest/West/ West Central Queens
Borough Park/ Southern/ Southwest/


Census data


Geology
N
L
M
I
K
G
H
E
F
C
D
A
B

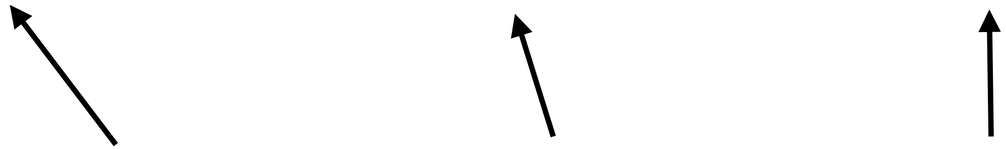

HARLEM
CENTRAL HARLEM
MORNINGSIDE HEIGHTS
SOUTH HARLEM
EAST HARLEM
UPPER WEST SIDE
CARNEGIE HILL
UPPER EAST SIDE
YORKVILLE
ASTORIA
EAST ELMHURST
JACKSON HEIGHTS
LONG ISLAND CITY
WOODSIDE
SUNNYSIDE
CORONA
FLUSHING
BAYSIDE
FRESH MEADOWS
MIDTOWN
TIMES SQUARE
MURRAY HILL
CHELSEA
WEST VILLAGE
UNION SQUARE
EAST VILLAGE
GREENPOINT
WILLIAMSBURG
EAST WILLIAMSBURG
SOHO
TRIBECA
CHINATOWN
LOWER EAST SIDE
FINANCIAL DISTRICT
DUMBO
DOWNTOWN BROOKLYN
BROOKLYN HEIGHTS
CLINTON HILL
BEDFORD-STUYVESANT
BUSHWICK
RIDGEWOOD
MIDDLE VILLAGE
FOREST HILLS
GLENDALE
KEW GARDENS
JAMAICA
QUEENS VILLAGE
CARROLL GARDENS
PROSPECT HEIGHTS
PARK SLOPE
GOWANUS
CROWN HEIGHTS
SOUTH SLOPE
PROSPECT LEFFERTS GARDENS
FLATBUSH
EAST FLATBUSH
CANARSIE
SUNSET PARK
BAY RIDGE
DYKER HEIGHTS
BENSONHURST
BOROUGH PARK
MIDWOOD
GRAVESEND
SHEEPSHEAD BAY
BRIGHTON BEACH
OCEAN PARKWAY
ROCKAWAY BEACH
ROCKAWAY PARK
BELLE HARBOR
FAR ROCKAWAY
STATEN ISLAND
PORT RICHMOND
WESTERLEIGH
STAPLETON
CLIFTON
ROSEBANK
GRASMERE
ARROCHAR
SOUTH BEACH
DONGAN HILLS
NEW DORP
NEW DORP BEACH
GREAT KILLS



(multi-detector &
multi-disciplinary)

(multimodal)

(self-supervised)



I want to construct a wide-ranging, multi-layered, data-driven representation of our Universe that can both complement and deepen our existing scientific understanding.

Illustration by Vivek Thakker

ANNALS OF ARTIFICIAL INTELLIGENCE

CHATGPT IS A BLURRY JPEG OF THE WEB

The AI industry has already established some effective methods for integrating text, images, video, speech, etc. across domains.

Will physics data present new challenges for these methods?

Physics data has some qualities that are relatively under-explored in mainstream ML.

Deep theoretical grounding → exploit known symmetries & simulations

Diverse data formats → custom vs. generic embeddings

Detector-specific biases → publish data without detector effects

Group anomalies → statistical anomaly detection

Highly multi-modal → unify inputs across an entire detector

Informed by universal physical laws → multi-detector/disciplinary analyses

Physics data has some qualities that are relatively under-explored in mainstream ML.

Deep theoretical grounding → exploit known symmetries & simulations

Diverse data formats → custom vs. generic embeddings

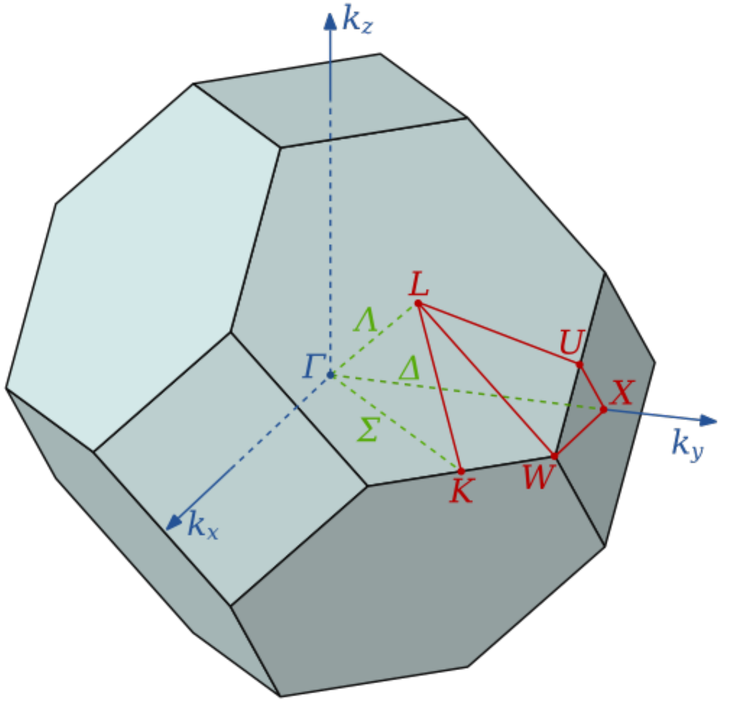
Detector-specific biases → publish data without detector effects

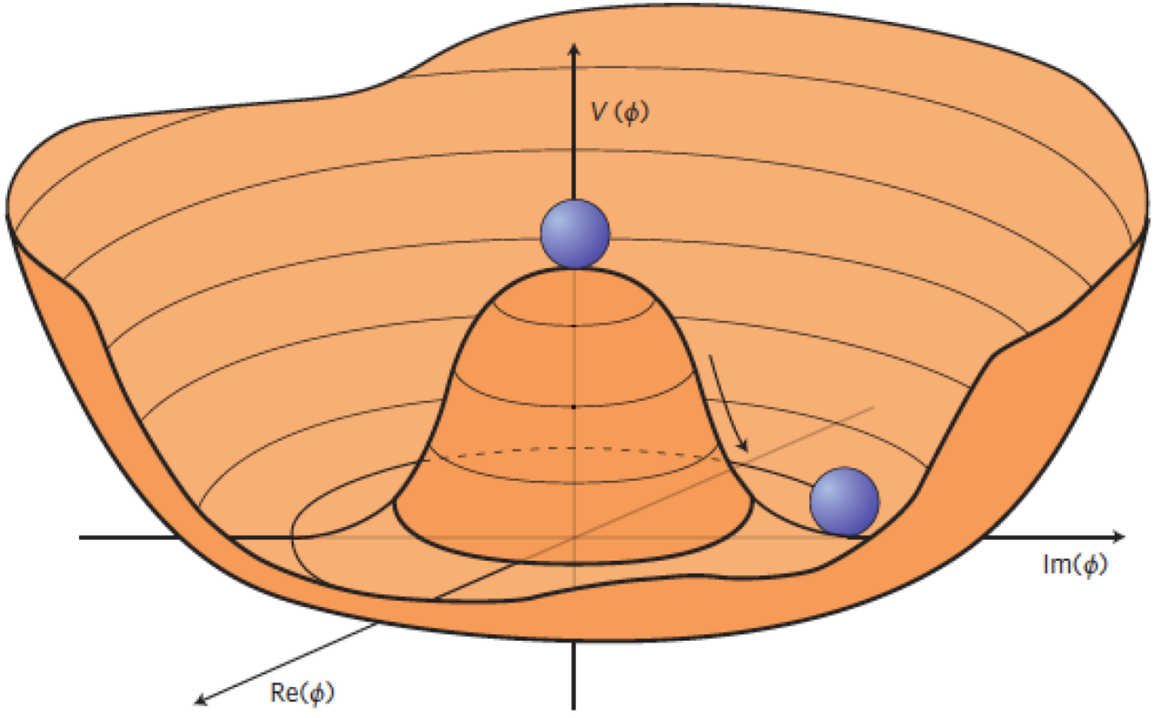
Group anomalies → statistical anomaly detection

Highly multi-modal → unify inputs across an entire detector

Informed by universal physical laws → multi-detector/disciplinary analyses

Unlike most fields, physics is grounded in deep & well-known mathematical frameworks.

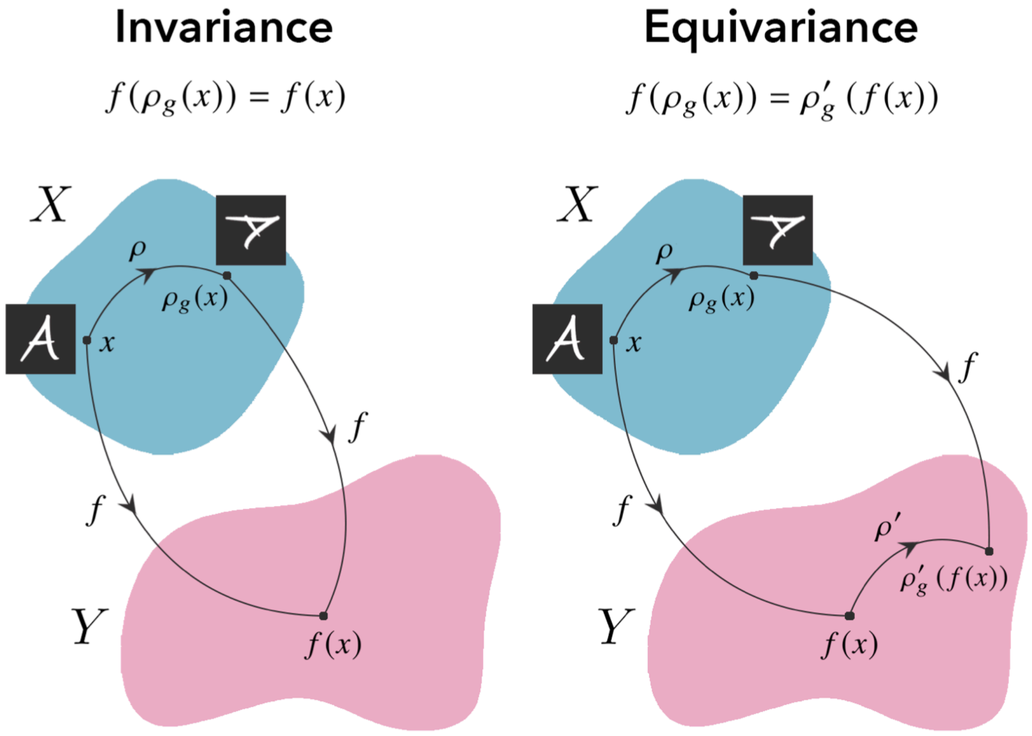


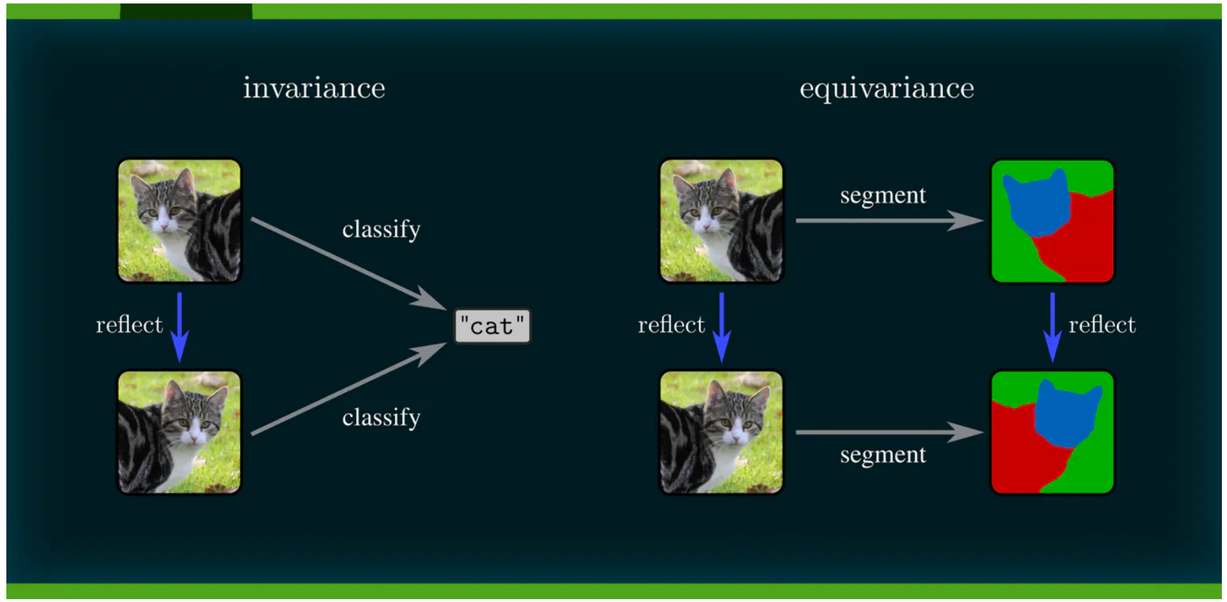


(particularly useful: symmetries!)

Unlike most fields, physics is grounded in deep & well-known mathematical frameworks.

Symmetries can be implicitly or explicitly encoded into neural networks:

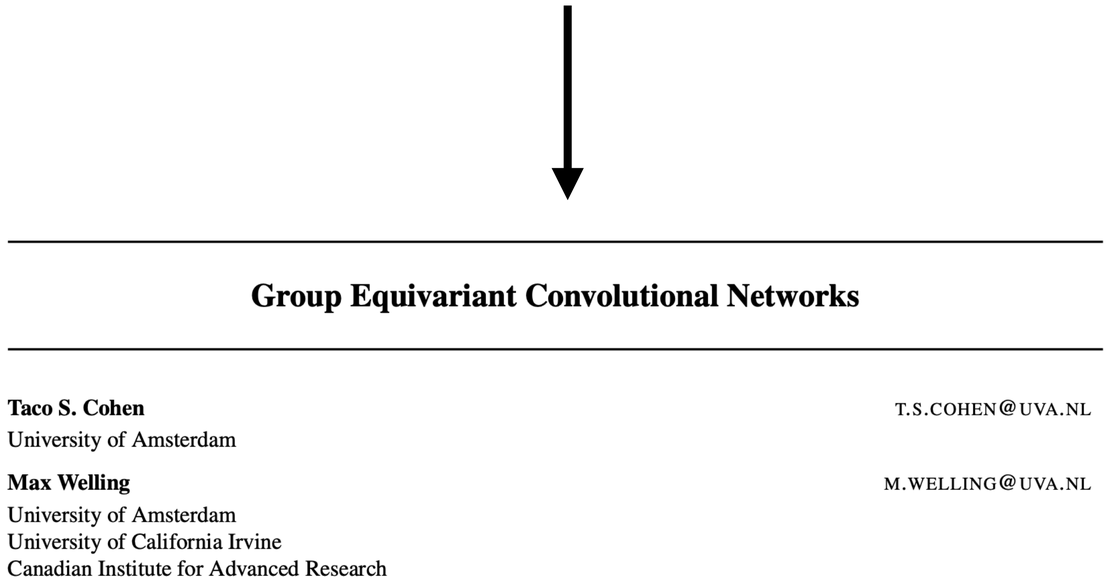




Unlike most fields, physics is grounded in deep & well-known mathematical frameworks.

We can do this even for much more complicated-looking symmetries:

$$\mathrm{SO}^+(1,3) \rtimes \mathbb{R}^{(1,3)} \times \mathrm{SU}(3)_\mathrm{C} \times \mathrm{SU}(2)_\mathrm{L} \times \mathrm{U}(1)_\mathrm{Y}$$

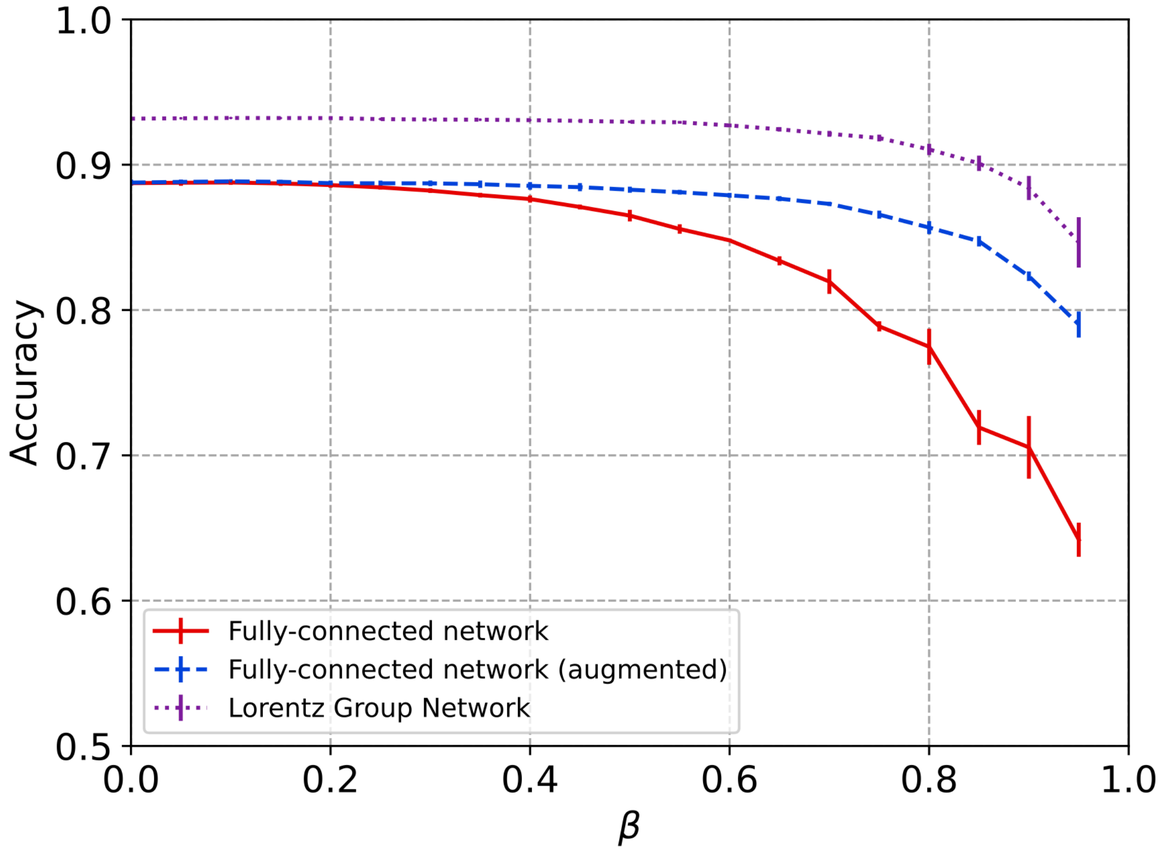


Group Equivariant Convolutional Networks

Taco S. Cohen T.S.COHEN@UVA.NL
University of Amsterdam

Max Welling M.WELLING@UVA.NL
University of Amsterdam
University of California Irvine
Canadian Institute for Advanced Research

Unlike most fields, physics is grounded in deep & well-known mathematical frameworks.



That said, the benefits of equivariant networks are most striking when data is highly limited — in a data-rich regime, the benefits are less clear.

Training %	Model	Accuracy	AUC	$1/\epsilon_B$
0.5%	ParticleNet	0.913	0.9687	199 ± 14
	LorentzNet	**0.932**	**0.9793**	$\mathbf{562 \pm 72}$
1%	ParticleNet	0.919	0.9734	287 ± 19
	LorentzNet	**0.932**	**0.9812**	$\mathbf{697 \pm 58}$
5%	ParticleNet	0.931	0.9807	609 ± 35
	LorentzNet	**0.937**	**0.9839**	$\mathbf{1108 \pm 84}$

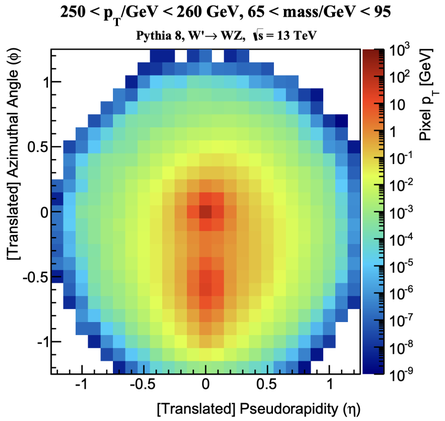
arXiv:2311.03094

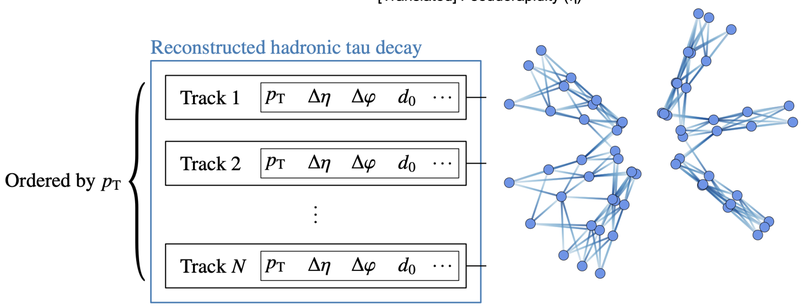
What level of structure is best for representing our data?

Prescribed structure

Learned structure

Industry-standard formats



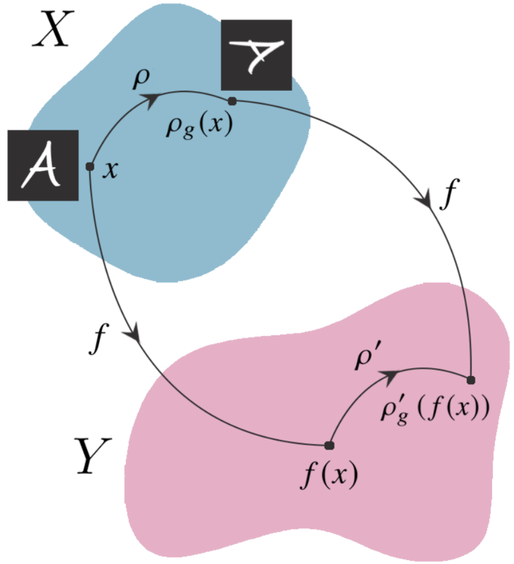


arXiv:1511.05190, arXiv:2007.13681

Physics-informed

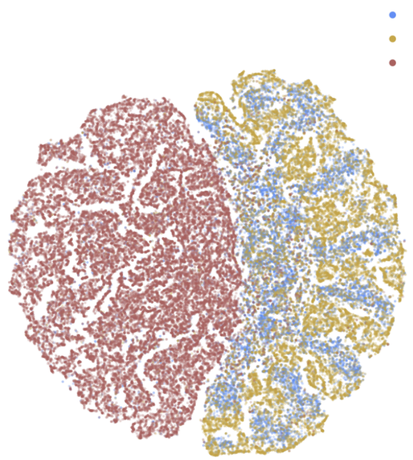
Equivariance

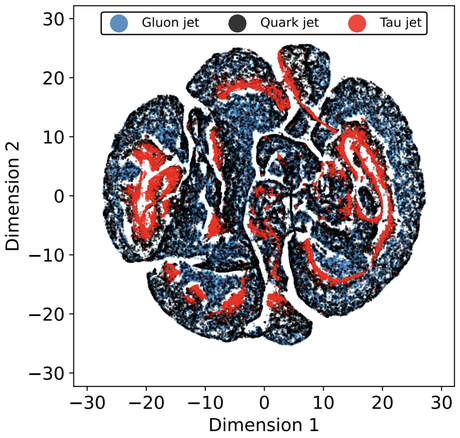
$$f(\rho_g(x)) = \rho'_g(f(x))$$



arXiv:2203.06153

Self-supervised



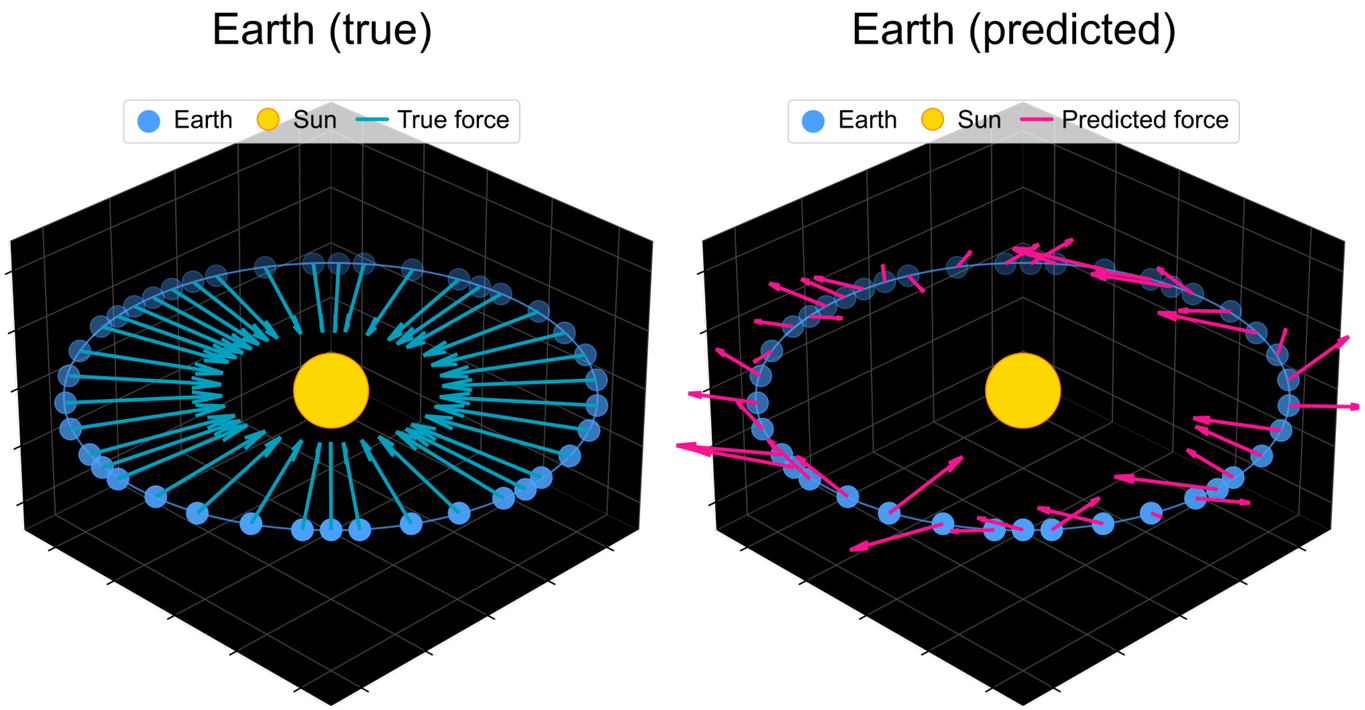


arXiv:2403.07066, arXiv:2503.11632

Unlike most fields, physics is grounded in deep & well-known mathematical frameworks.

Another benefit: we often have high-quality simulations!

But also: a rigorous standard for interpretability (maybe)



Physics data has some qualities that are relatively under-explored in mainstream ML.

Deep theoretical grounding → exploit known symmetries & simulations

Diverse data formats → custom vs. generic embeddings

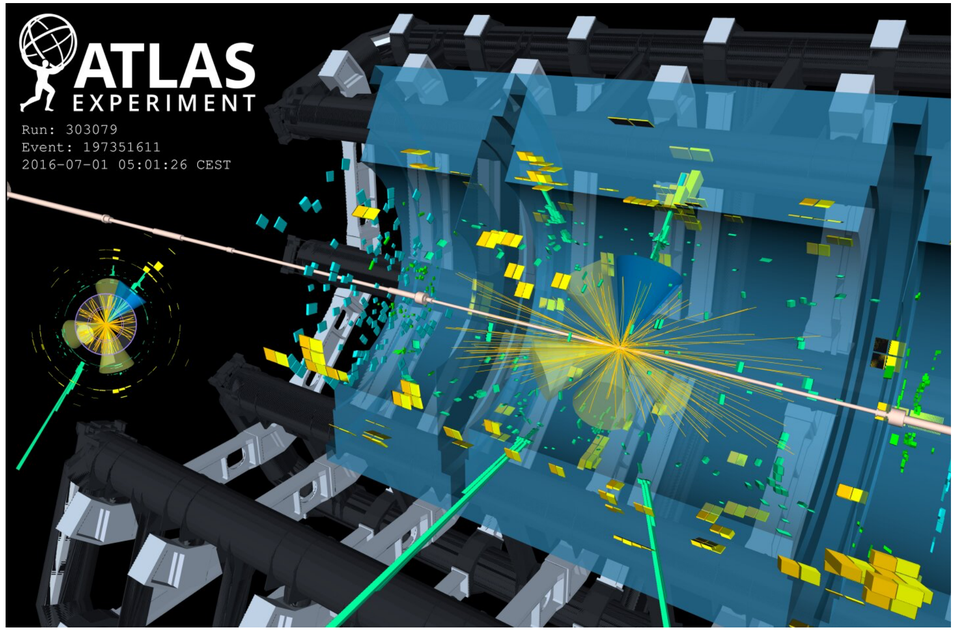
Detector-specific biases → publish data without detector effects

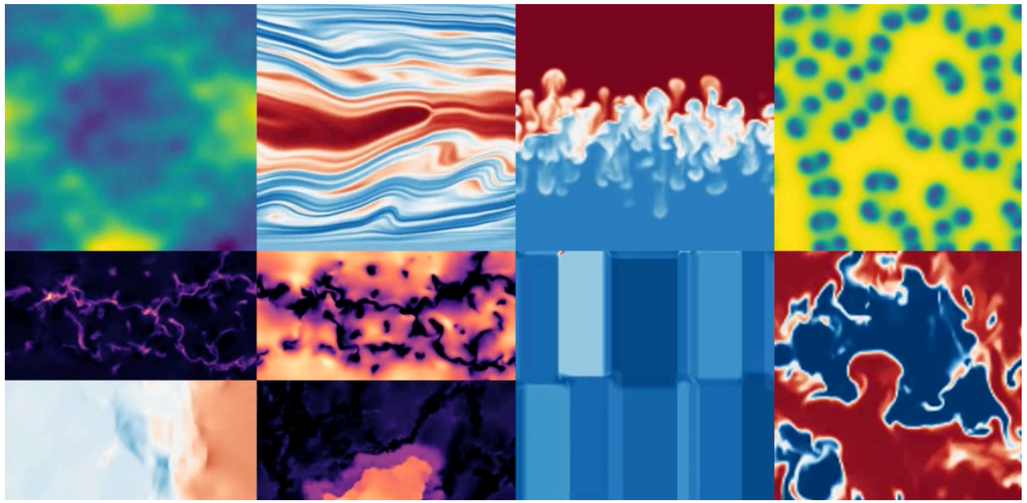
Group anomalies → statistical anomaly detection

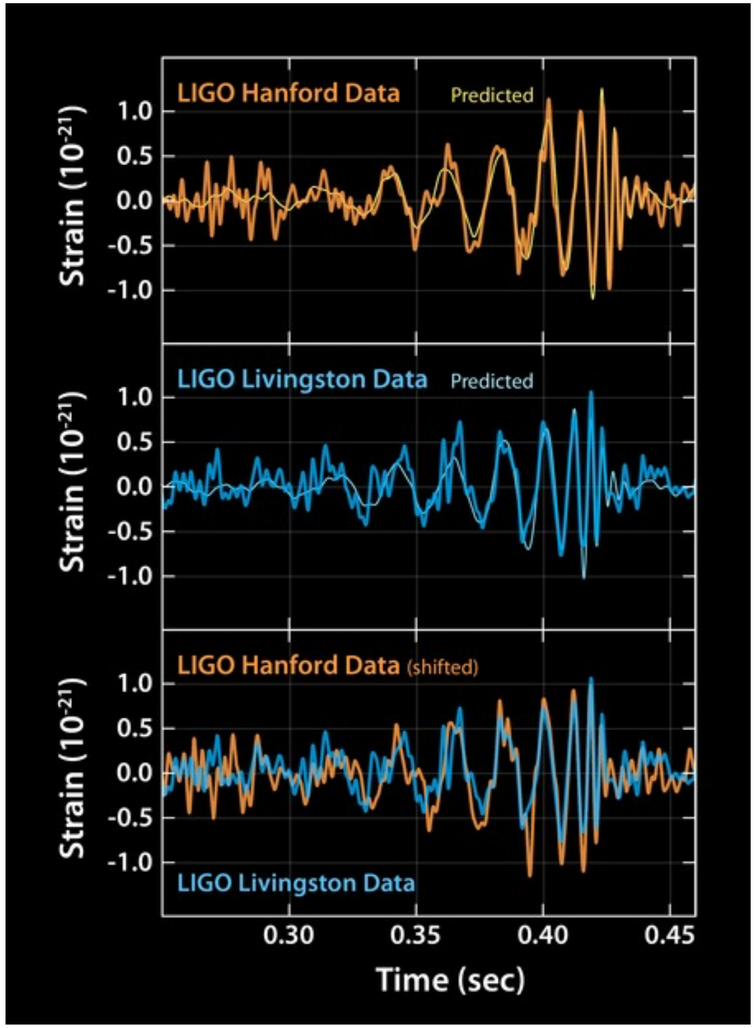
Highly multi-modal → unify inputs across an entire detector

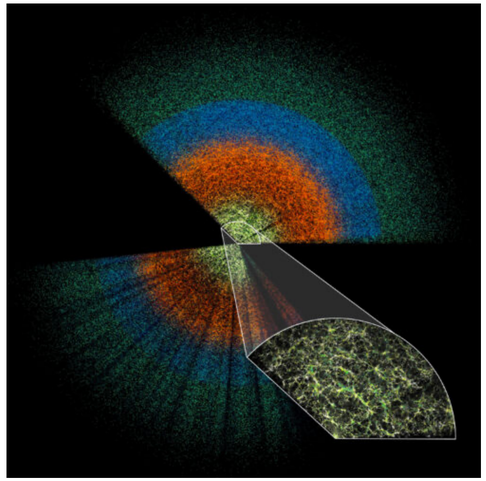
Informed by universal physical laws → multi-detector/disciplinary analyses

Physics data comes in all kinds of custom formats that are difficult to natively combine.



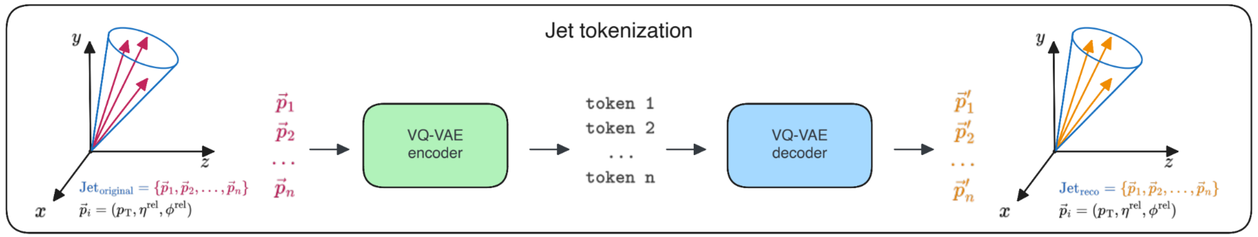




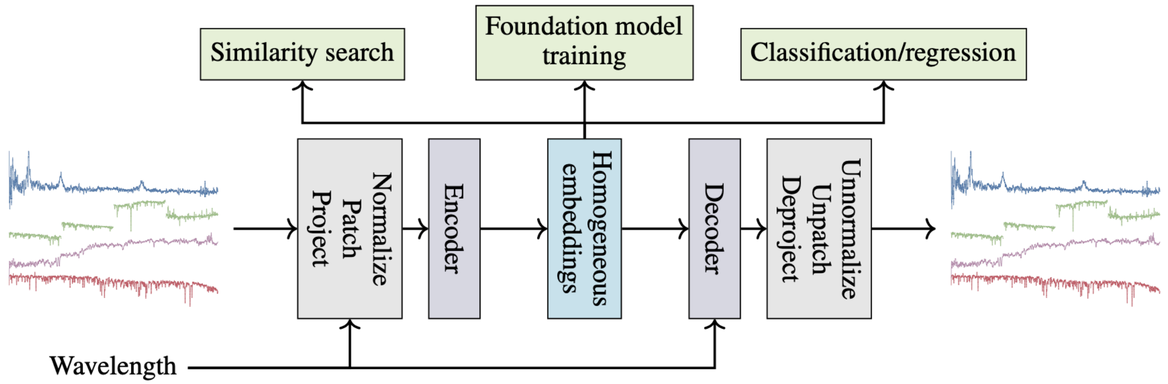


Physics data comes in all kinds of custom formats that are difficult to natively combine.

Custom embeddings:
precise, but potentially tedious



arXiv:2403.05618



arXiv:2510.17959

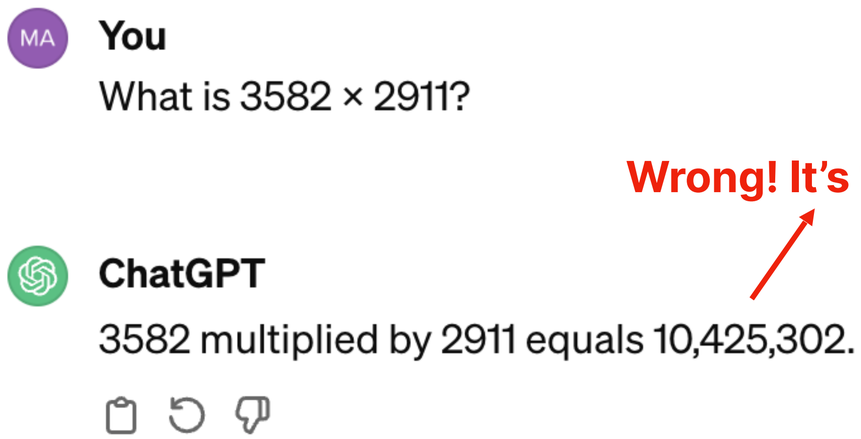
Generic embedding?

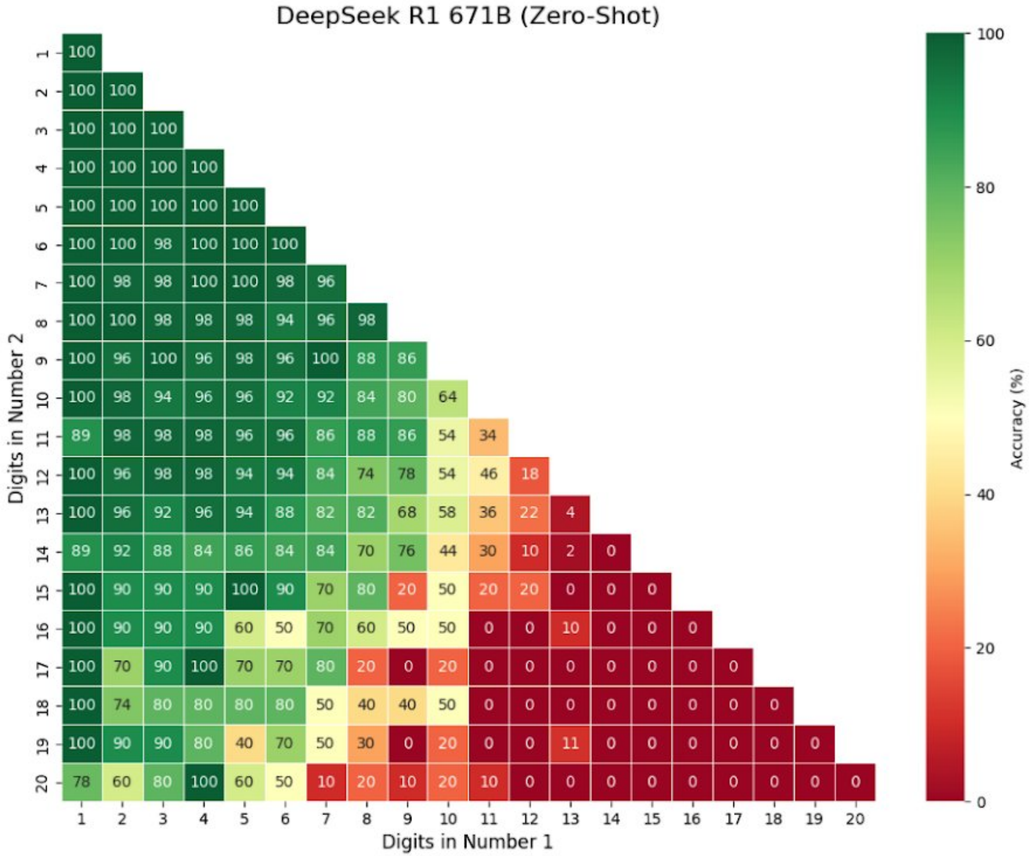
```
{'description':{
  'planet0':{
    'm':4.842,'a':2.679,'e':0.923
    },
  'planet1':{
    'm':4.813,'a':1.0,'e':0.191
    },
  'stepsize': 0.8},
'data':[
  [[2.525,-0.397],[0.963,-0.232]],
  [[2.557,0.112],[0.830,0.542]],
  [[2.483,0.618],[0.1832,0.981]],
  [[2.318,1.100],[-0.577,0.823]],
  [[2.066,1.541],[-0.994,0.173]]
  ]}
```

Physics data comes in all kinds of custom formats that are difficult to natively combine.

Rendering data as text is easy to combine, but most language tokenizers hinder mathematical abilities.

e.g. GPT-4 fails at 4-digit multiplication
over 95% of the time without tool use:





arXiv:2305.18654

Physics data comes in all kinds of custom formats that are difficult to natively combine.

LLMs struggle to understand what makes numbers distinct from other language-based tokens.

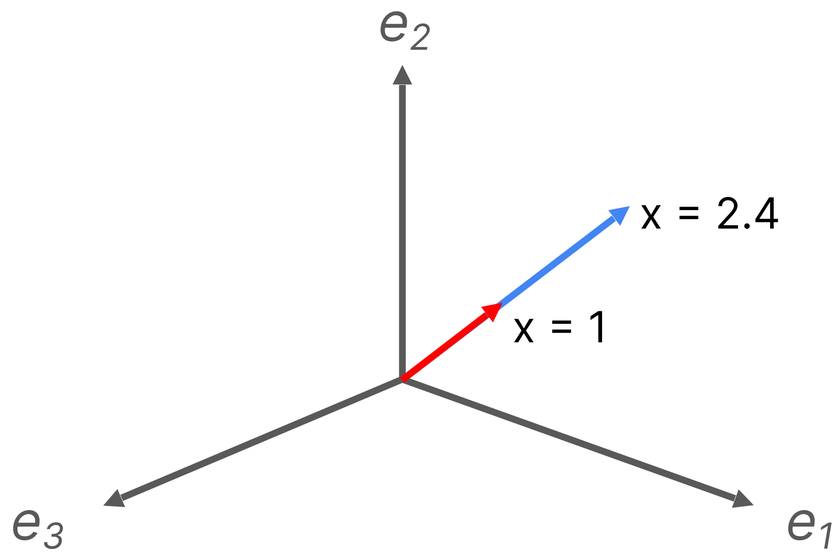
Moreover, they struggle to represent numbers with a finite dictionary of tokens:

Encoding	3.14	$-6.02.10^{23}$	Tokens / coefficient	Size of vocabulary
P10	`[+, 3, 1, 4, E-2]`	`[-, 6, 0, 2, E21]`	5	210
P1000	`[+, 314, E-2]`	`[-, 602, E21]`	3	1100
B1999	`[314, E-2]`	`[-602, E21]`	2	2000
FP15	`[FP314/-2]`	`[FP-602/21]`	1	30000

Physics data comes in all kinds of custom formats that are difficult to natively combine.

One possibility is to add a dedicated numerical token with a magnitude that scales with the number itself:

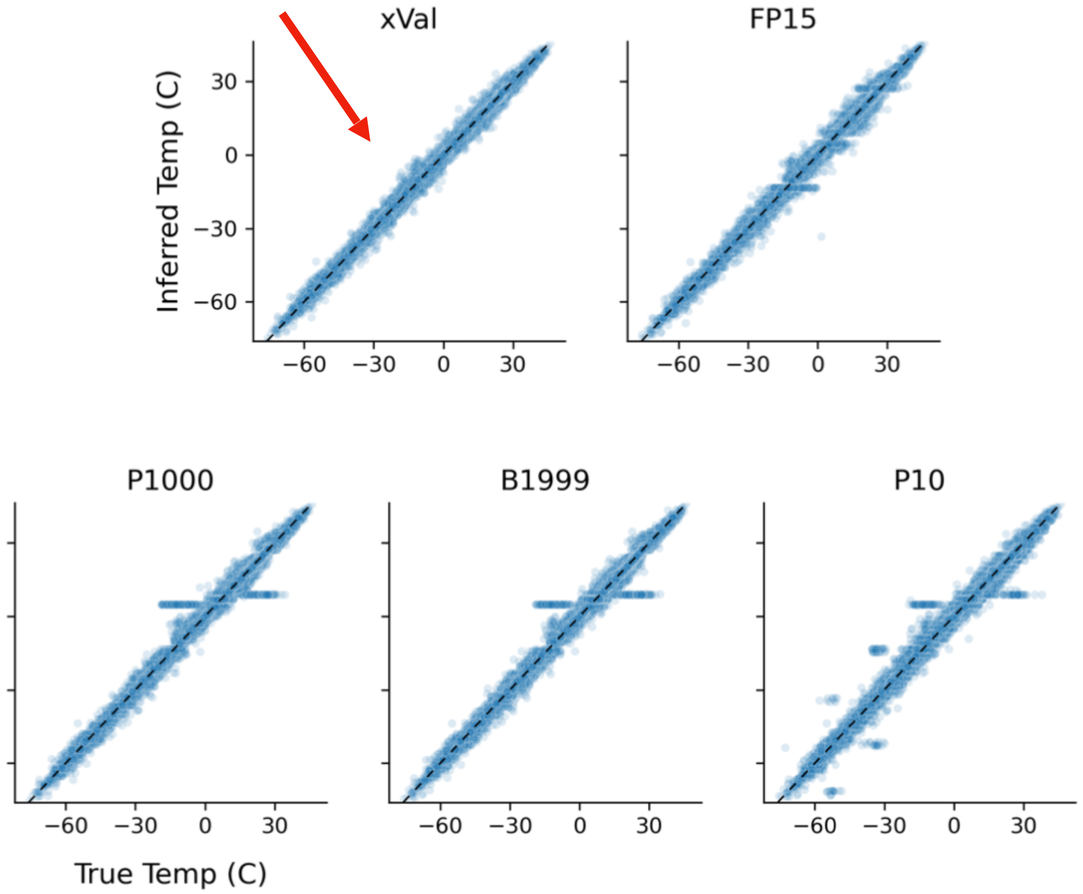
+



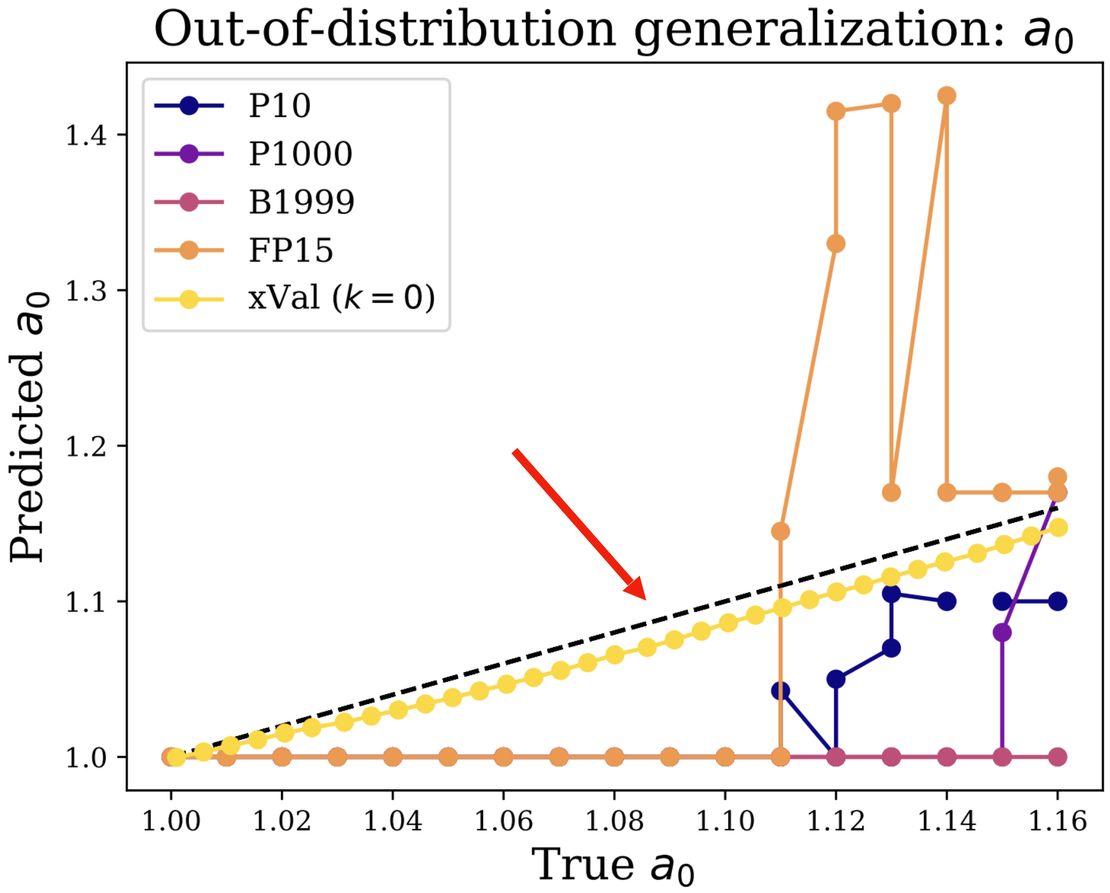
arXiv:2310.02989

Dedicated numerical encodings can help enable rendering physics data as language.

In-distribution

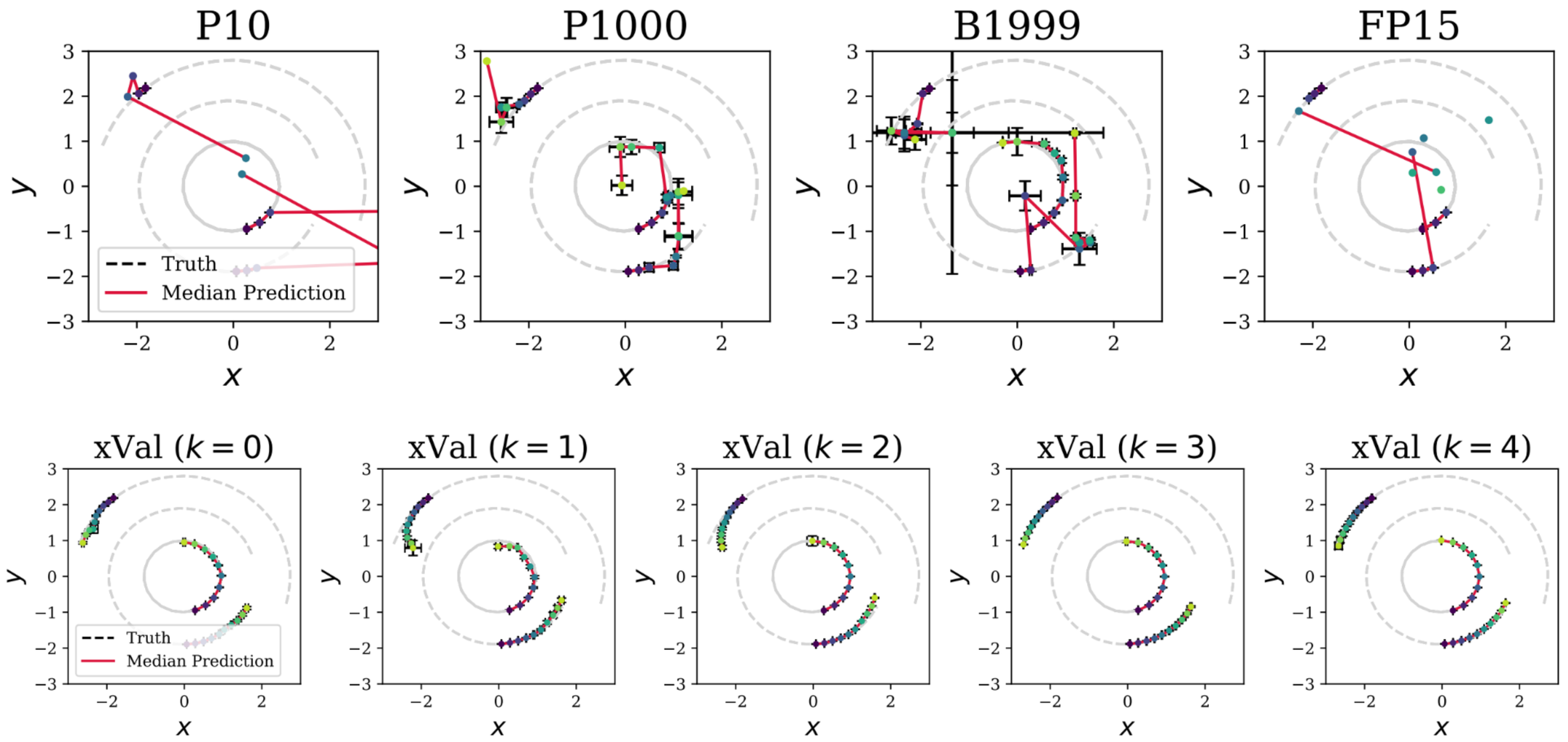


Out-of-distribution

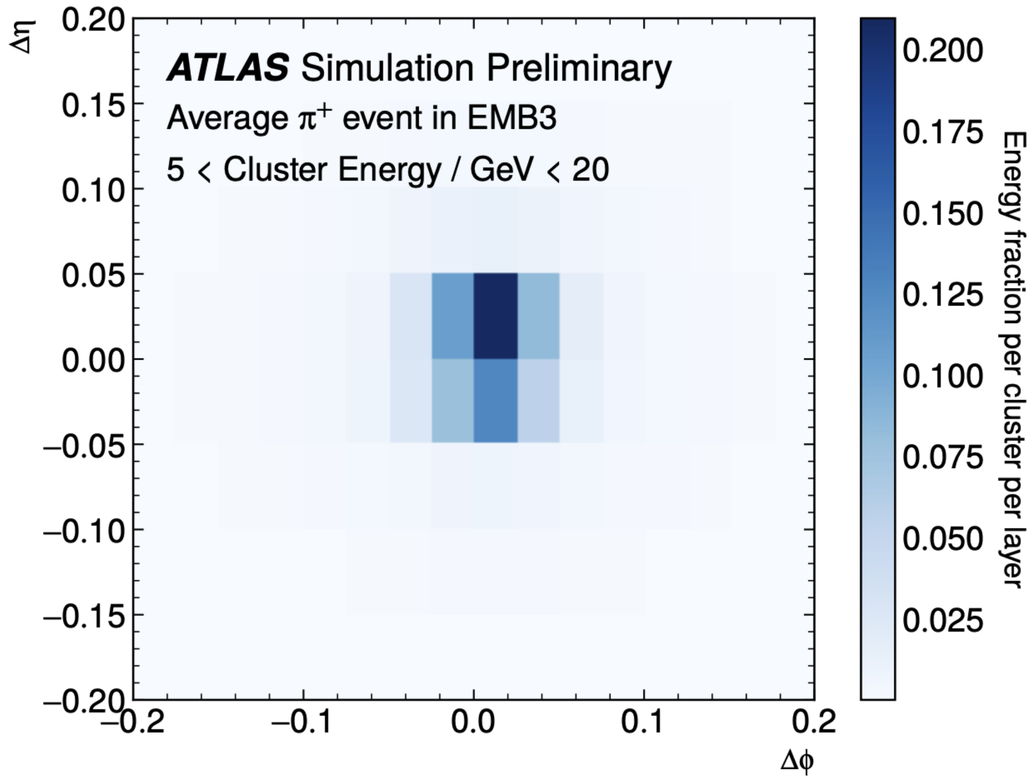


Dedicated numerical encodings can help enable rendering physics data as language.

This can even predict time series data, despite not being trained autoregressively:



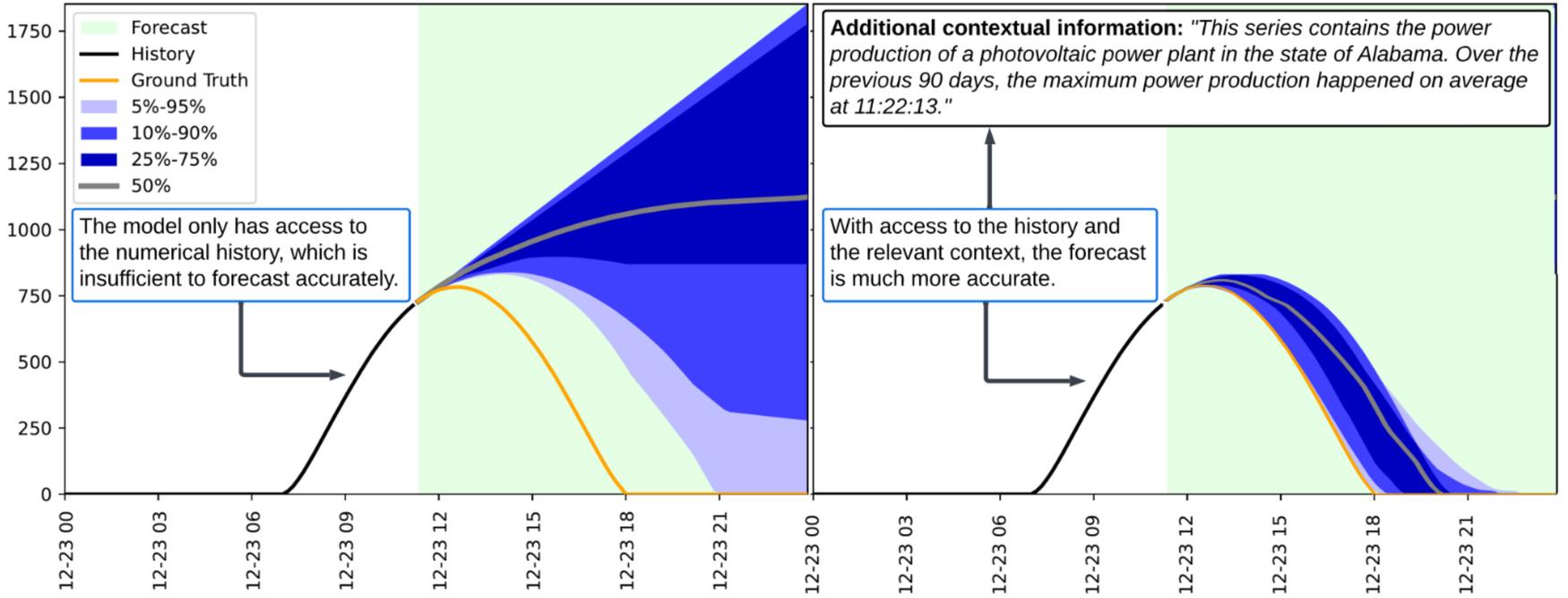
Moreover, physics data often needs context — perhaps we should treat text as a critical modality:



Moreover, physics data often needs context — perhaps we should treat text as a critical modality:

Context is Key: A Benchmark for Forecasting with Essential Textual Information

*Andrew R. Williams [1 2 3] *Arjun Ashok [1 2 3]
†Étienne Marcotte [1] †Valentina Zantedeschi [1 4] Jithendaraa Subramanian [1 2 5] Roland Riachi [2]
James Requeima [6] Alexandre Lacoste [1] Irina Rish [2 3] †Nicolas Chapados [1 2 7] †Alexandre Drouin [1 2 4]



arXiv:2410.18959

Physics data has some qualities that are relatively under-explored in mainstream ML.

Deep theoretical grounding → exploit known symmetries & simulations

Diverse data formats → custom vs. generic embeddings

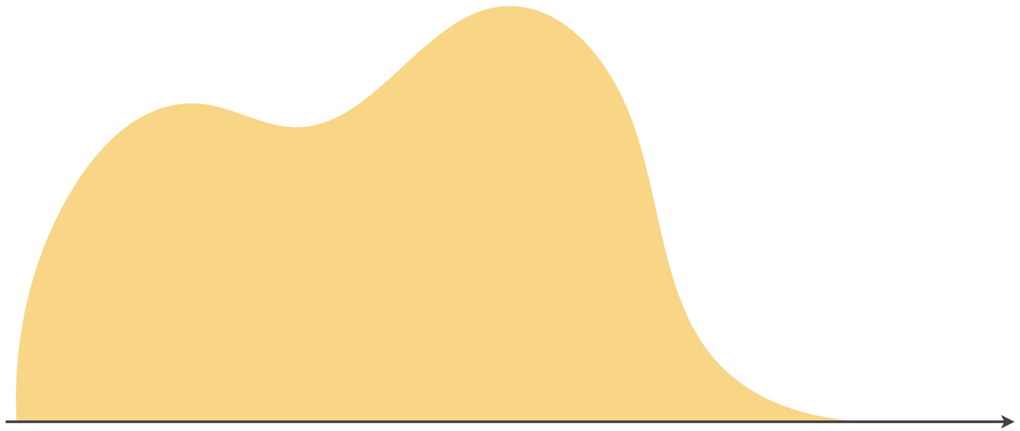
Detector-specific biases → publish data without detector effects

Group anomalies → statistical anomaly detection

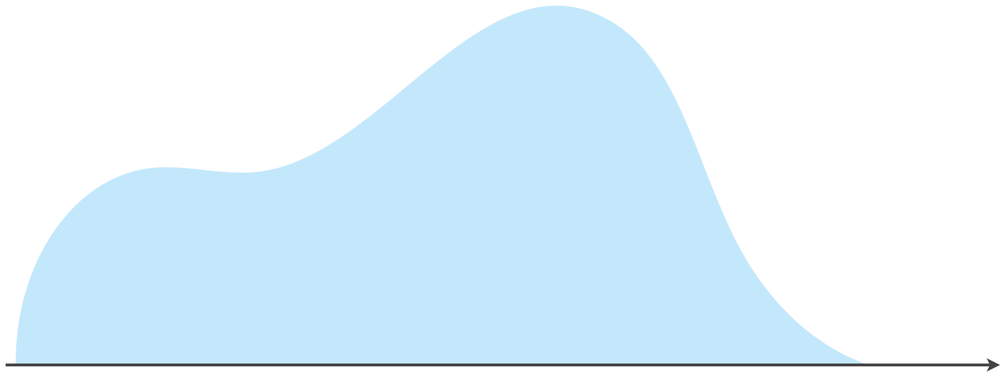
Highly multi-modal → unify inputs across an entire detector

Informed by universal physical laws → multi-detector/disciplinary analyses

High-dimensional & unpinned unfolding allows us to publish more flexible datasets.



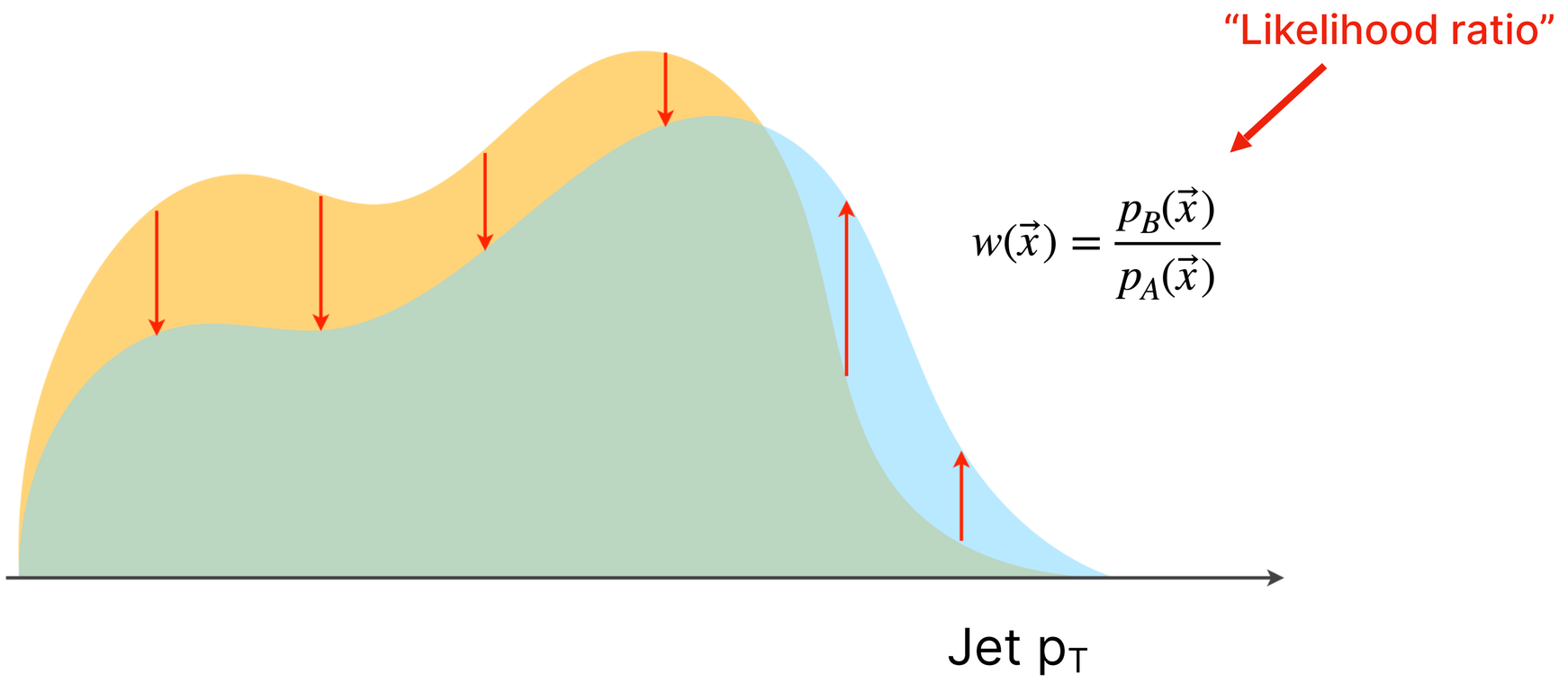
Measured jet p_T



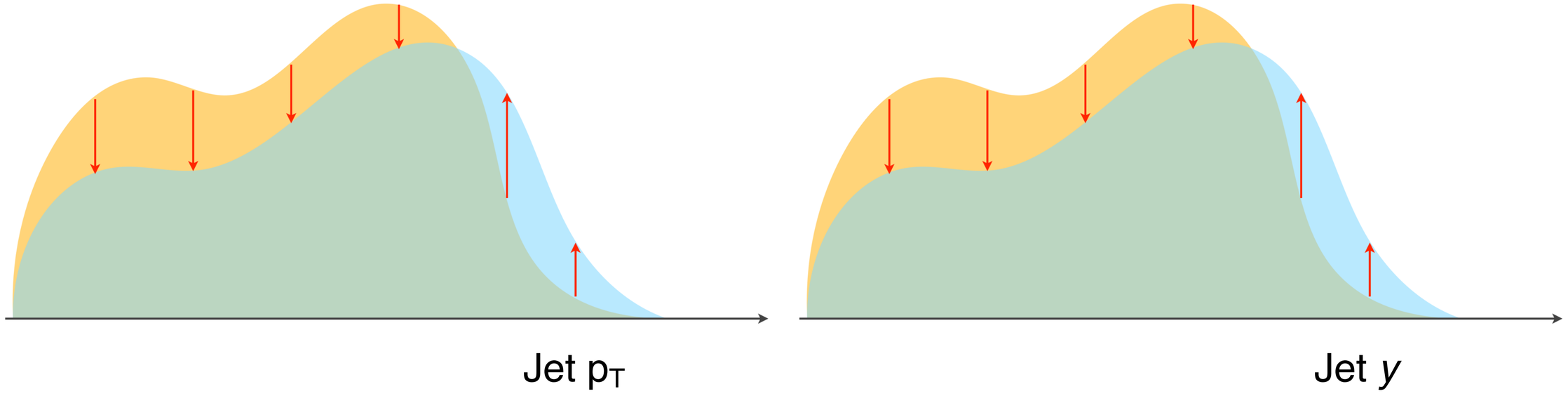
True jet p_T

High-dimensional & unpinned unfolding allows us to publish more flexible datasets.

Learns a high-dimensional reweighting function based on the ratio of the probability densities:



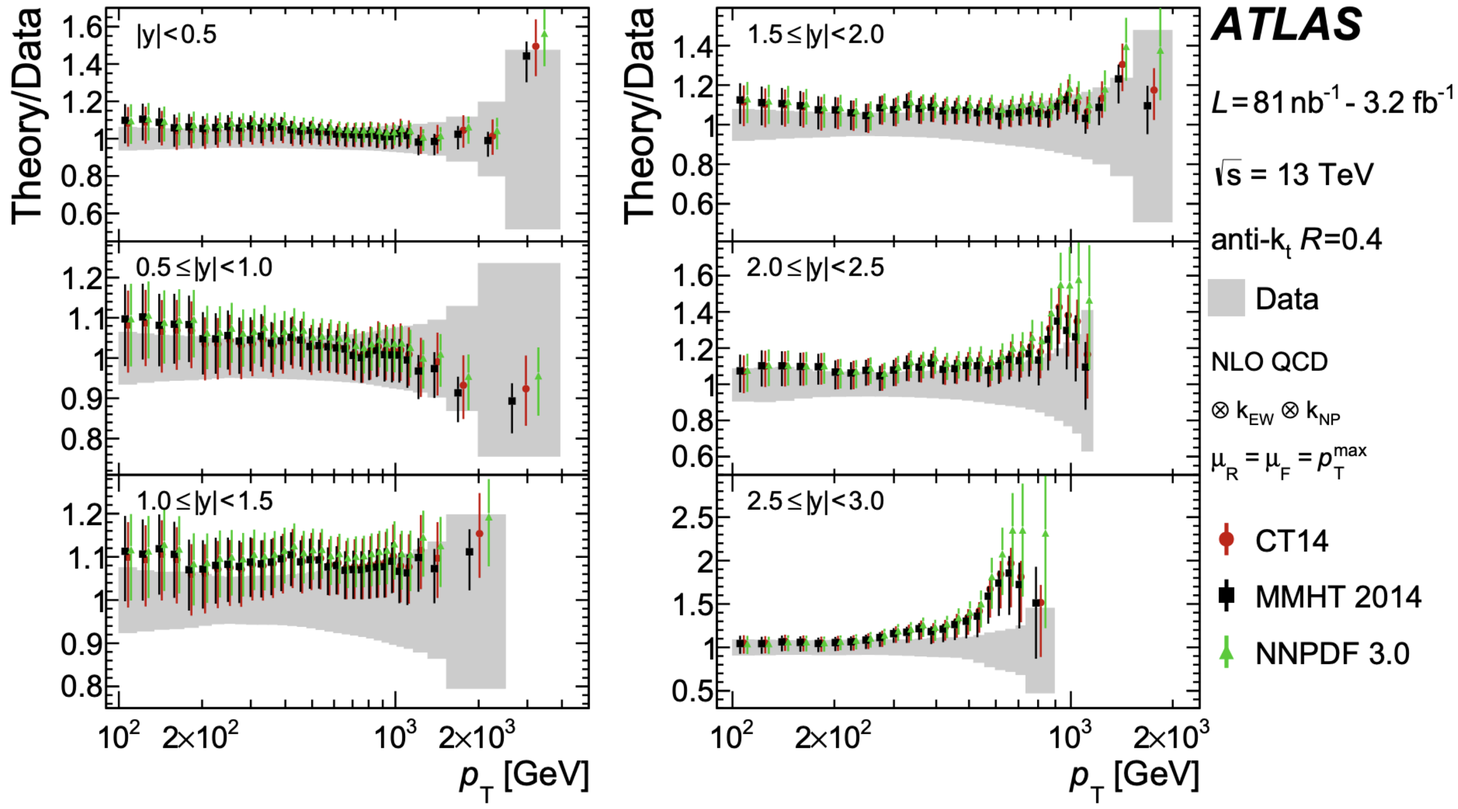
High-dimensional & unpinned unfolding allows us to publish more flexible datasets.



High-dimensional & unpinned unfolding allows us to publish more flexible datasets.

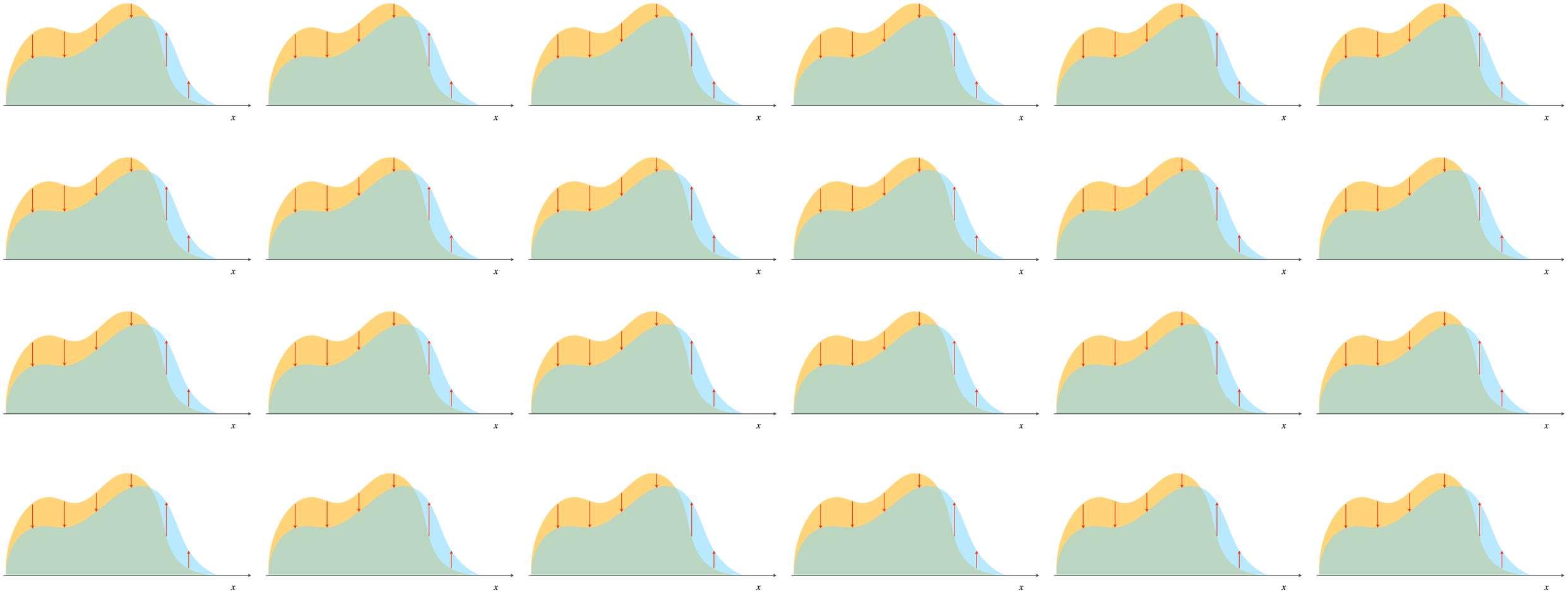
Many LHC measurements are published as binned differential cross-sections at the particle level.

(So far from the LHC, for only one or two observables at a time).

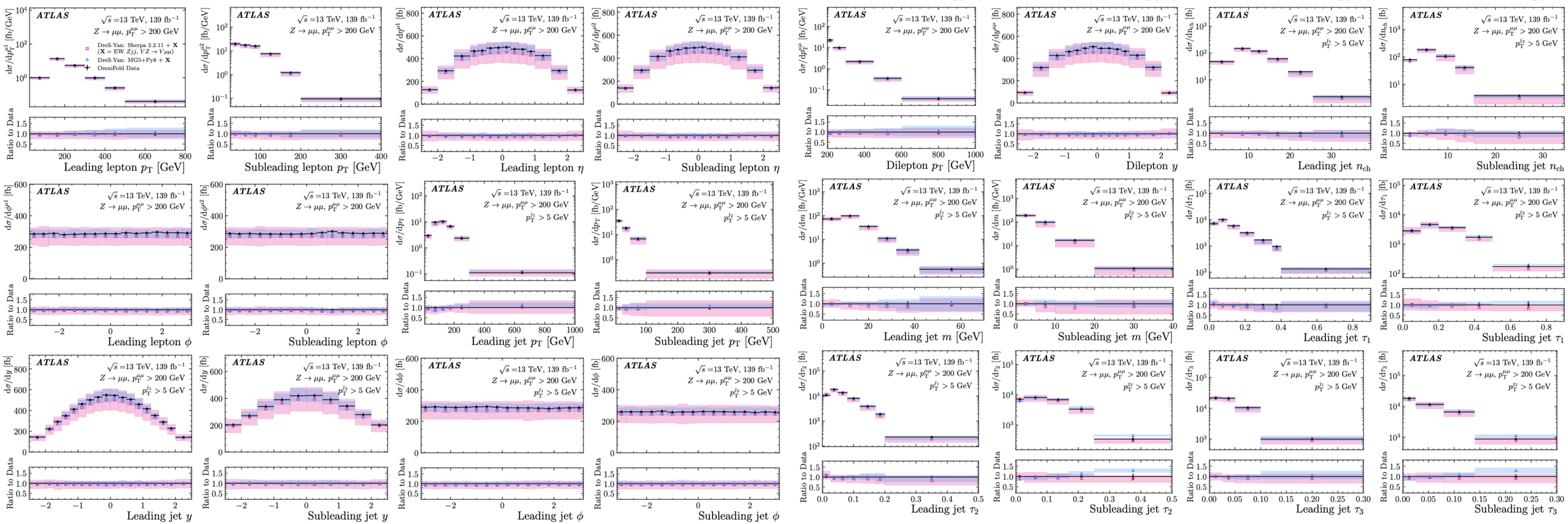


arXiv:1711.02692 • CERN-EP-2017-157

High-dimensional & unpinned unfolding allows us to publish more flexible datasets.



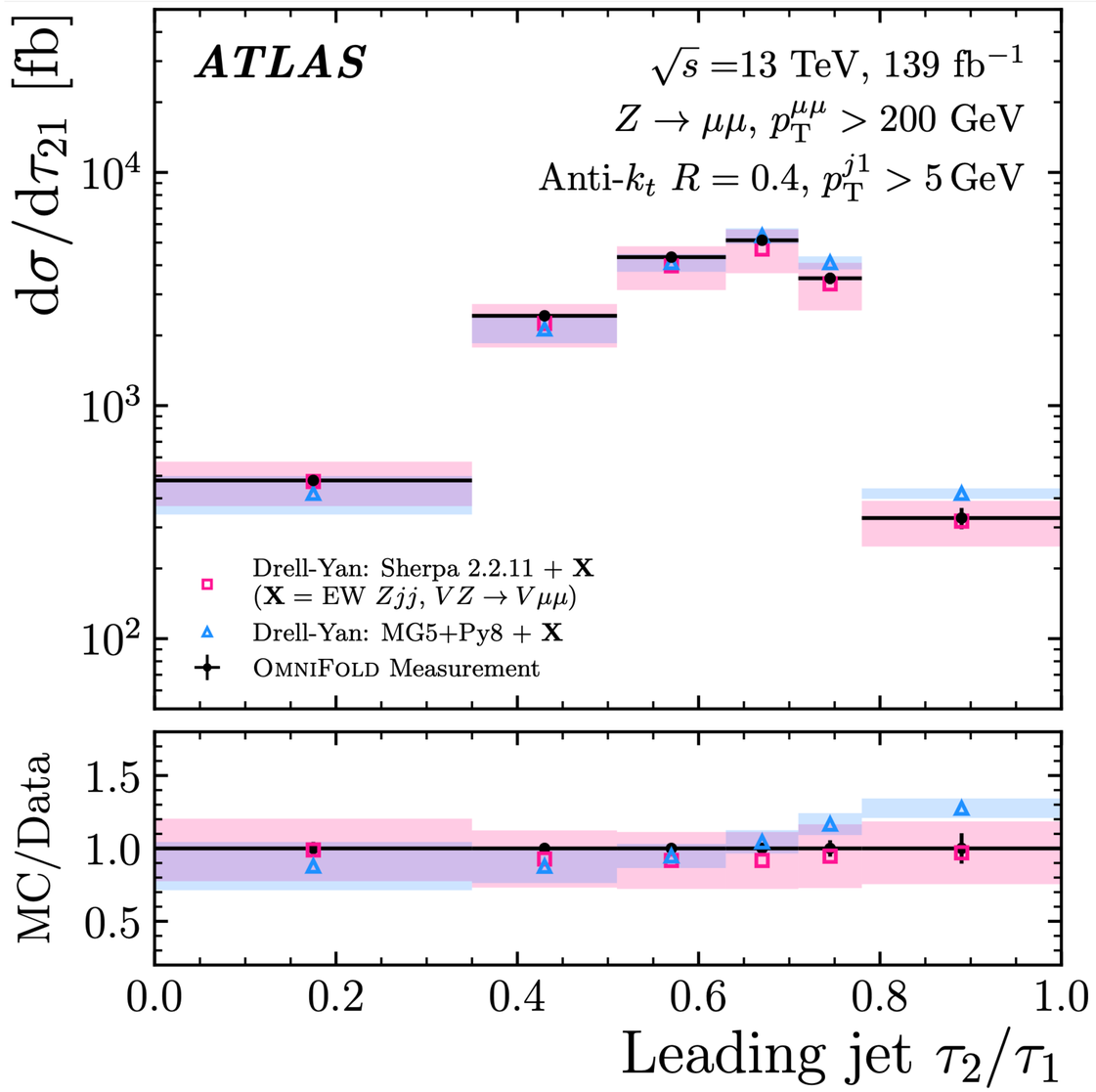
High-dimensional & unpinned unfolding allows us to publish more flexible datasets.



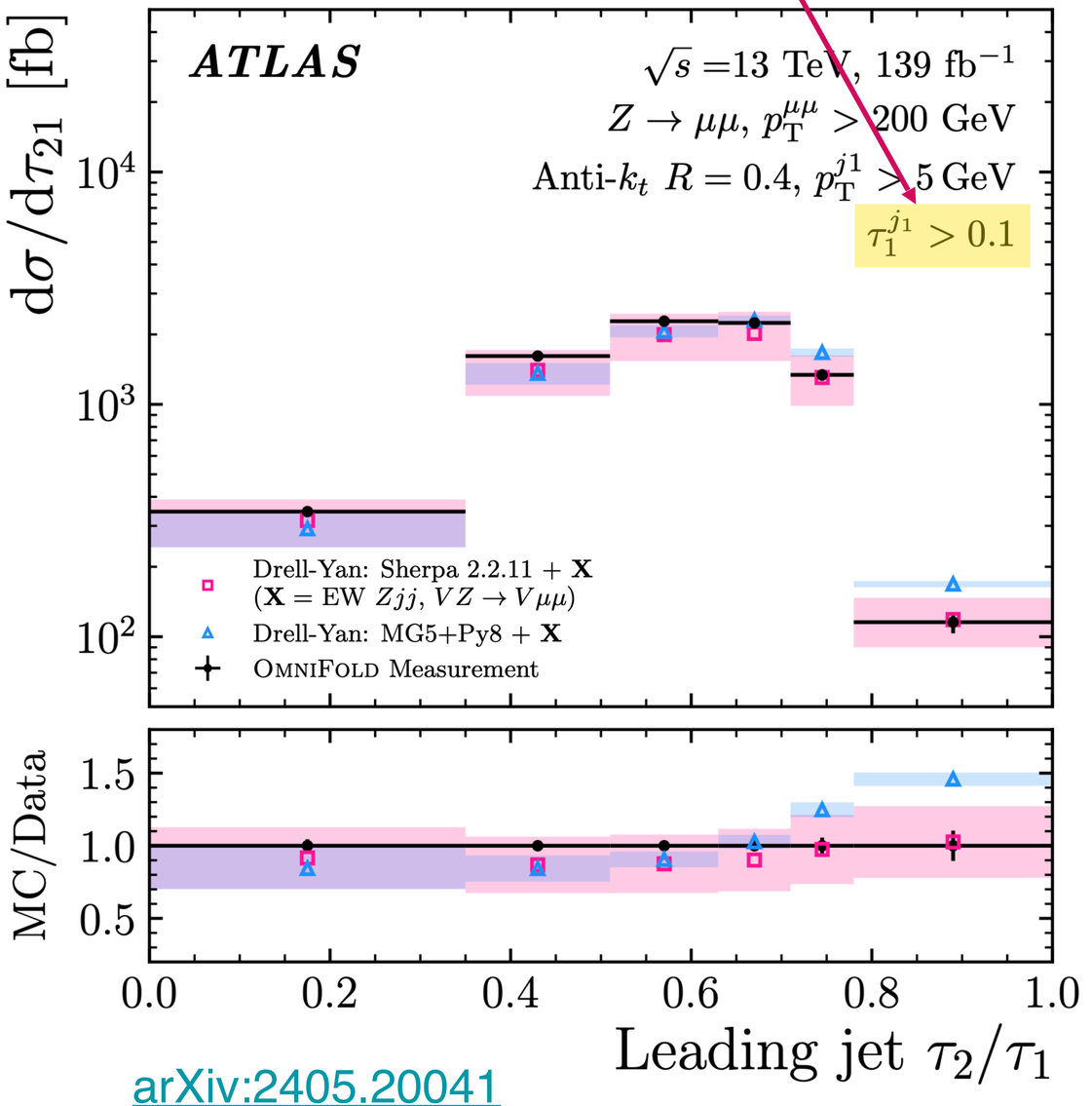
arXiv:2405.20041

Let's construct some new observables…

$$\tau_{21} = \tau_2/\tau_1$$



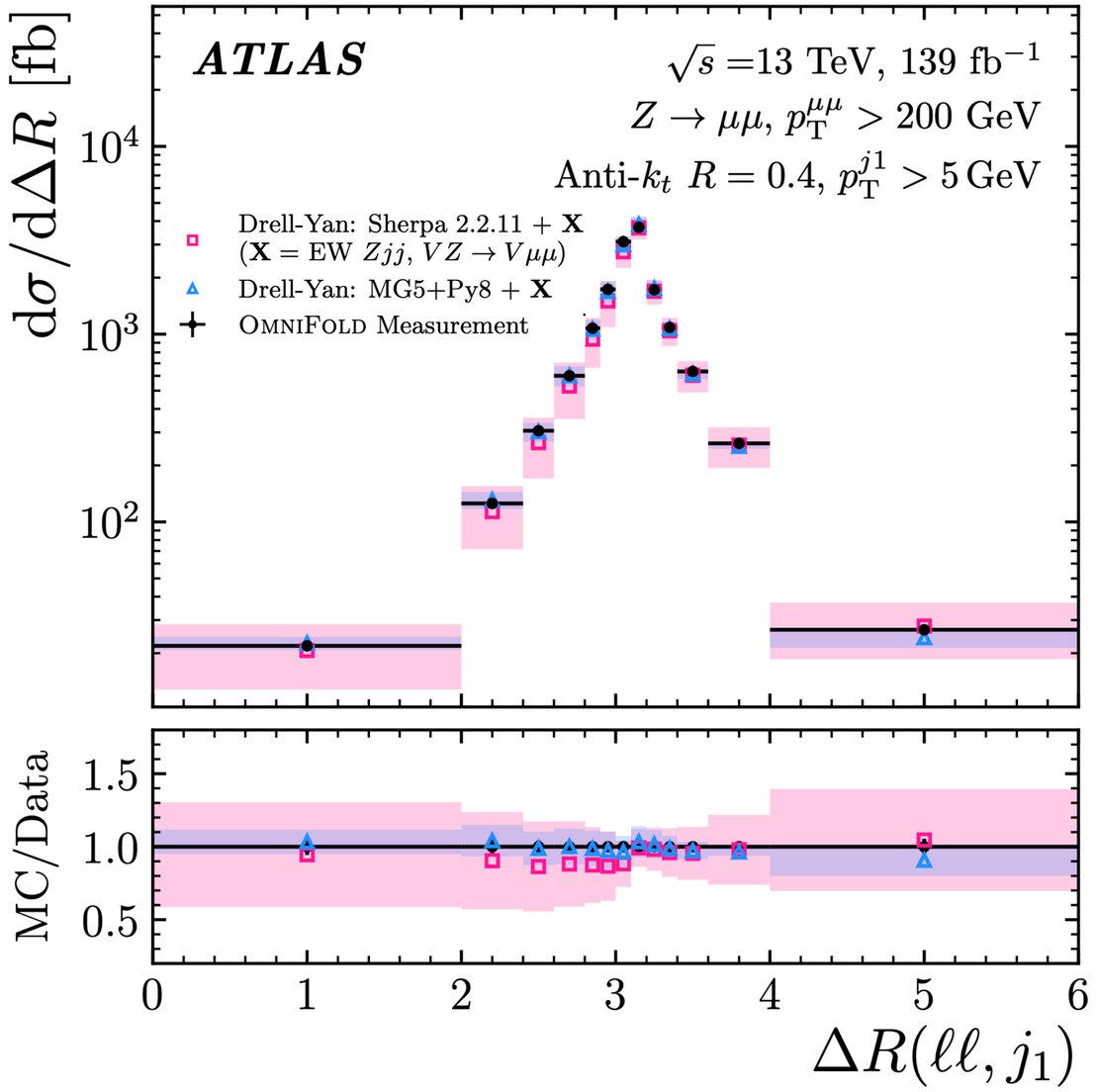
IRC-safe phase space!



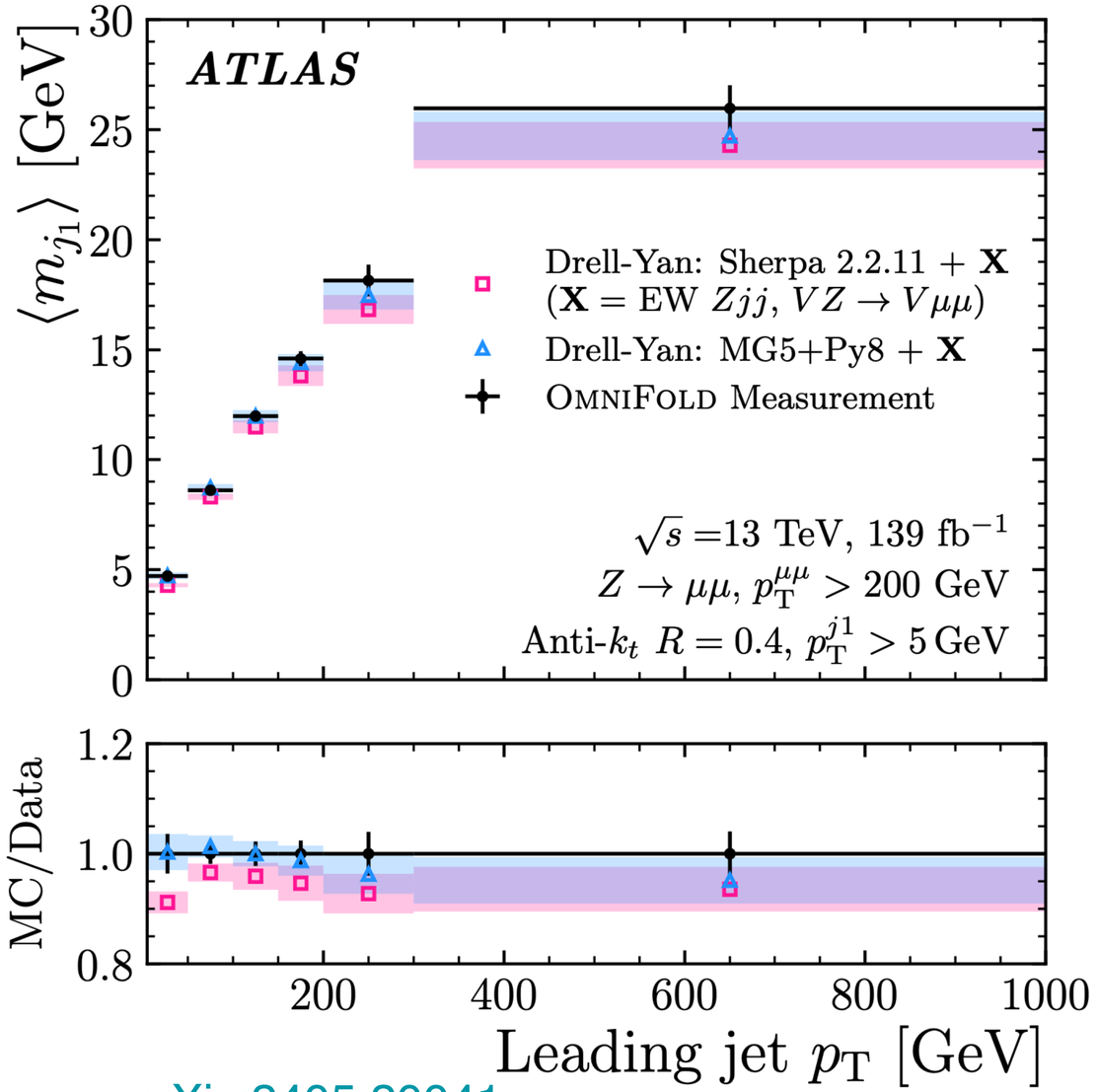
arXiv:2405.20041

Let’s construct some new observables…

$\Delta R(\ell\ell, j_1)$



Average m_{j1} in bins of leading jet p_T



arXiv:2405.20041

High-dimensional & unpinned unfolding allows us to publish more flexible datasets.

	pT_ll	pT_l1	pT_l2	eta_l1	...	weights_trackPtScale	weights_theoryPSjet	weights_theoryPSsoft
0	479.442780	288.466919	198.183929	-0.117443	...	0.003174	0.002844	0.003195
1	274.524994	166.120789	125.378044	0.313321	...	0.008168	0.008563	0.008236
2	462.713226	335.697479	133.157684	0.766387	...	0.001638	0.001724	0.001890
3	215.157608	189.518021	25.711994	1.083798	...	0.004669	0.004622	0.004648
4	222.458313	128.850159	108.589226	-0.635713	...	0.002102	0.002417	0.002129
...	...	...	...	...	...	...	...	...
418009	934.971924	738.464722	196.525192	0.102944	...	0.000069	0.000070	0.000061
418010	245.813461	166.847061	93.757919	1.308837	...	0.000193	0.000189	0.000203
418011	478.670349	378.737518	108.016479	-0.328871	...	0.001969	0.001813	0.001825
418012	278.586029	249.255356	43.581135	0.632484	...	0.003238	0.003101	0.003090
418013	244.505249	219.796280	40.357105	1.833223	...	0.000947	0.000968	0.000957

High-dimensional & unpinned unfolding allows us to publish more flexible datasets.

A Practical Guide to Unbinned Unfolding

Florencia Canelli,[1,*] Kyle Cormier,[1,*] Andrew Cudd,[2,†] Dag Gillberg,[3,‡] Roger G. Huang,[4,†] Weijie Jin,[1,*] Sookhyun Lee,[5,§] Vinicius Mikuni,[4,6,¶] Laura Miller,[7] Benjamin Nachman,[4,8,9,**] Jingjing Pan,[4,10,**] Tanmay Pani,[11,††] Mariel Pettee,[4,‡] Youqi Song,[10,††] and Fernando Torales Acosta[4,¶]

Experiment	Paper Link	Dimensions	Final State	Momentum Selection
ATLAS [23]	2405.20041	24	Z+jets	$p_{\mathrm{T}}^{\ell\ell} > 200$ GeV
ATLAS [24]	2502.02062	6	Dijets	$p_{\mathrm{T}}^{j1} > 240$ GeV & $p_{\mathrm{T}}^{j1} < 1.5\, p_{\mathrm{T}}^{j2}$
CMS [25]	2505.17850	8	Minimum bias	> 2 charged particles with $p_{\mathrm{T}} > 0.5$ GeV
H1 [26]	2108.12376	8*	High Q^2 DIS	$Q^2 > 150$ GeV^2
H1 [27]	2303.13620	10	High Q^2 DIS	$Q^2 > 150$ GeV^2
H1 [28]	2412.14092	8*	High Q^2 DIS	$Q^2 > 150$ GeV^2
H1 [29]	H1prelim-25-031	Variable	High Q^2 DIS	$Q^2 > 150$ GeV^2
LHCb [30]	2208.11691	4	Z+hadrons in jets	$20 < p_{\mathrm{T}}^{j} < 100$ GeV and $p_{\mathrm{T}}^{h} > 0.25$ GeV
STAR [31]	2307.07718	6	Jets	$20 < p_{\mathrm{T}}^{j} < 50$ GeV
STAR [32]	2403.13921	7	Jets in heavy ions	$20 < p_{\mathrm{T}}^{j} < 45$ GeV
T2K [33]	2504.06857	6	Muon + Proton	$p^{p} > 450$ MeV for single transverse variables

2507.09582

Physics data has some qualities that are relatively under-explored in mainstream ML.

Deep theoretical grounding → exploit known symmetries & simulations

Diverse data formats → custom vs. generic embeddings

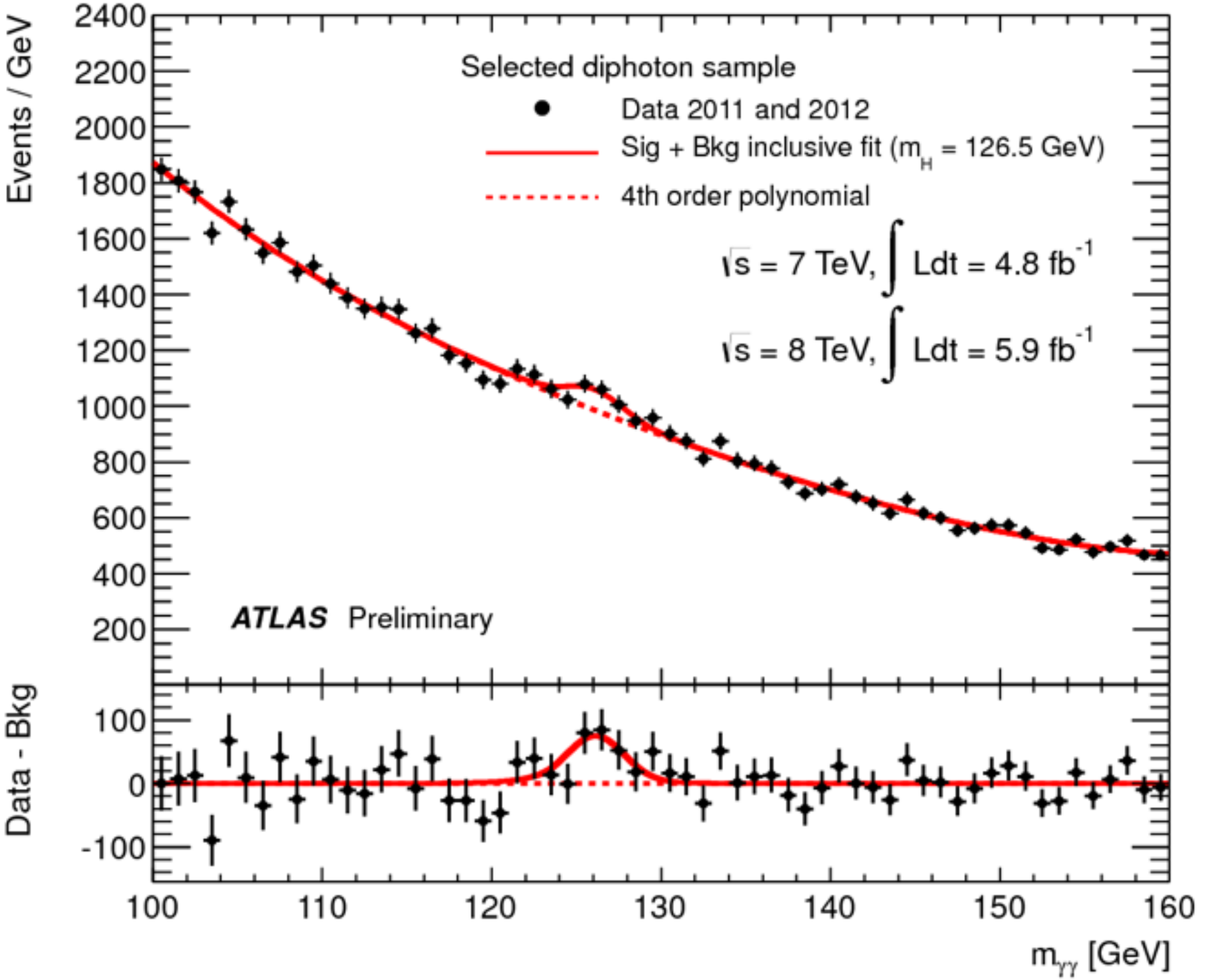
Detector-specific biases → publish data without detector effects

Group anomalies → statistical anomaly detection

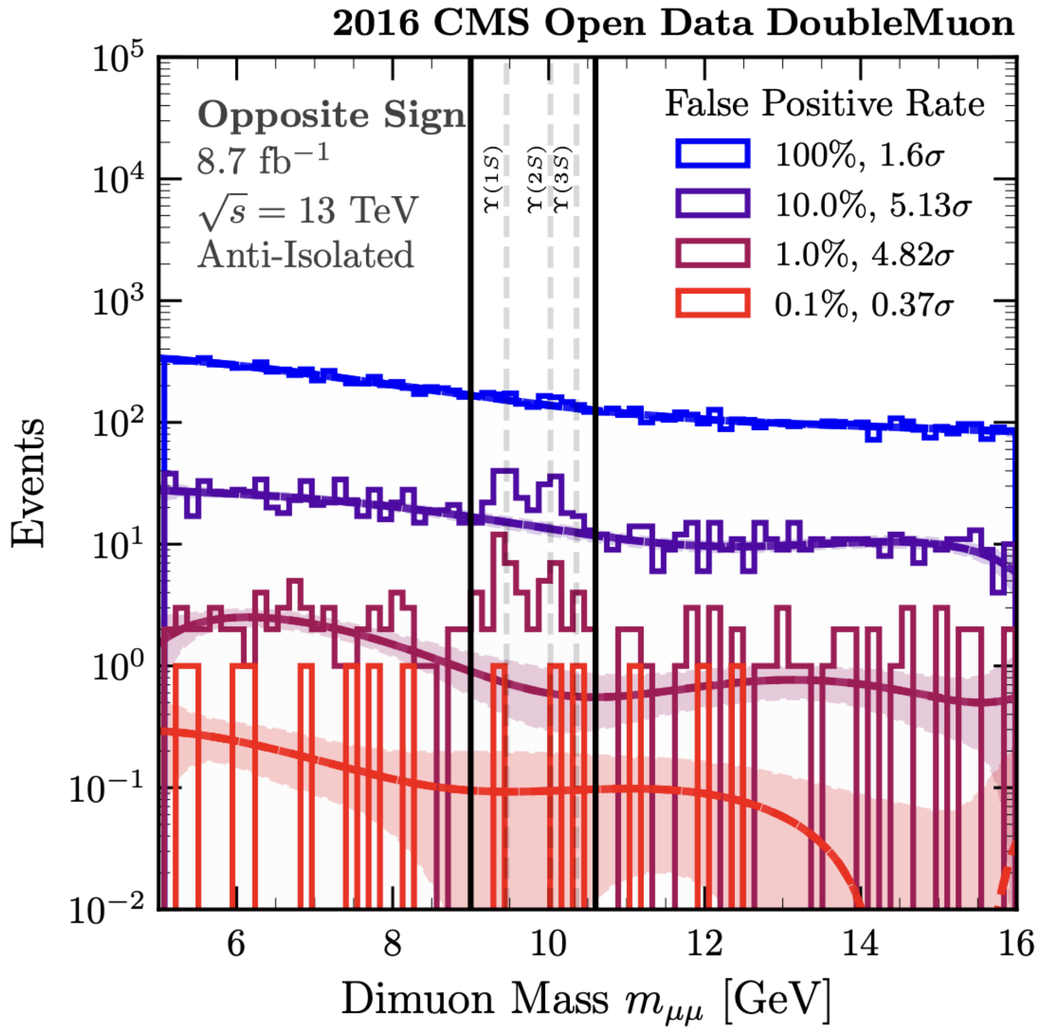
Highly multi-modal → unify inputs across an entire detector

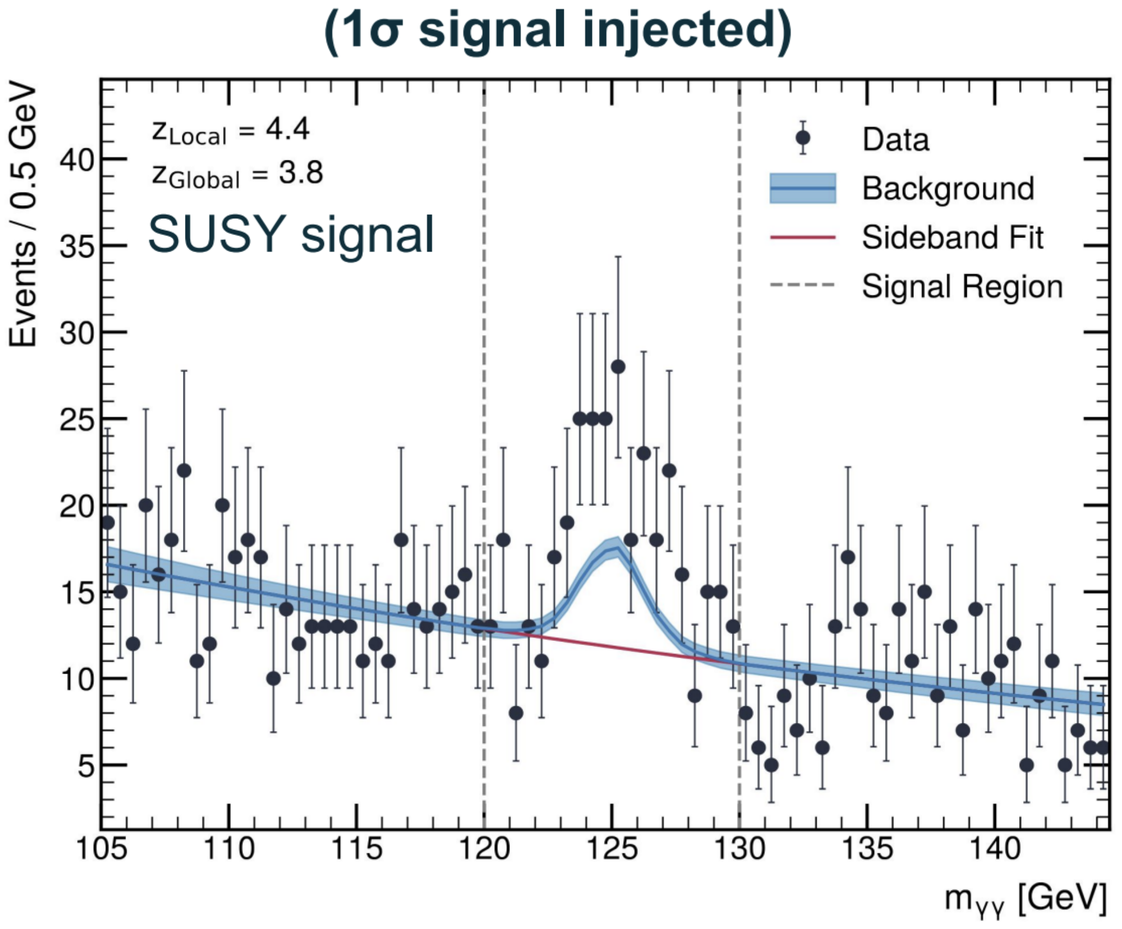
Informed by universal physical laws → multi-detector/disciplinary analyses

Anomaly detection in particle physics targets group anomalies, not single-event outliers.



Anomaly detection in particle physics targets group anomalies, not single-event outliers.



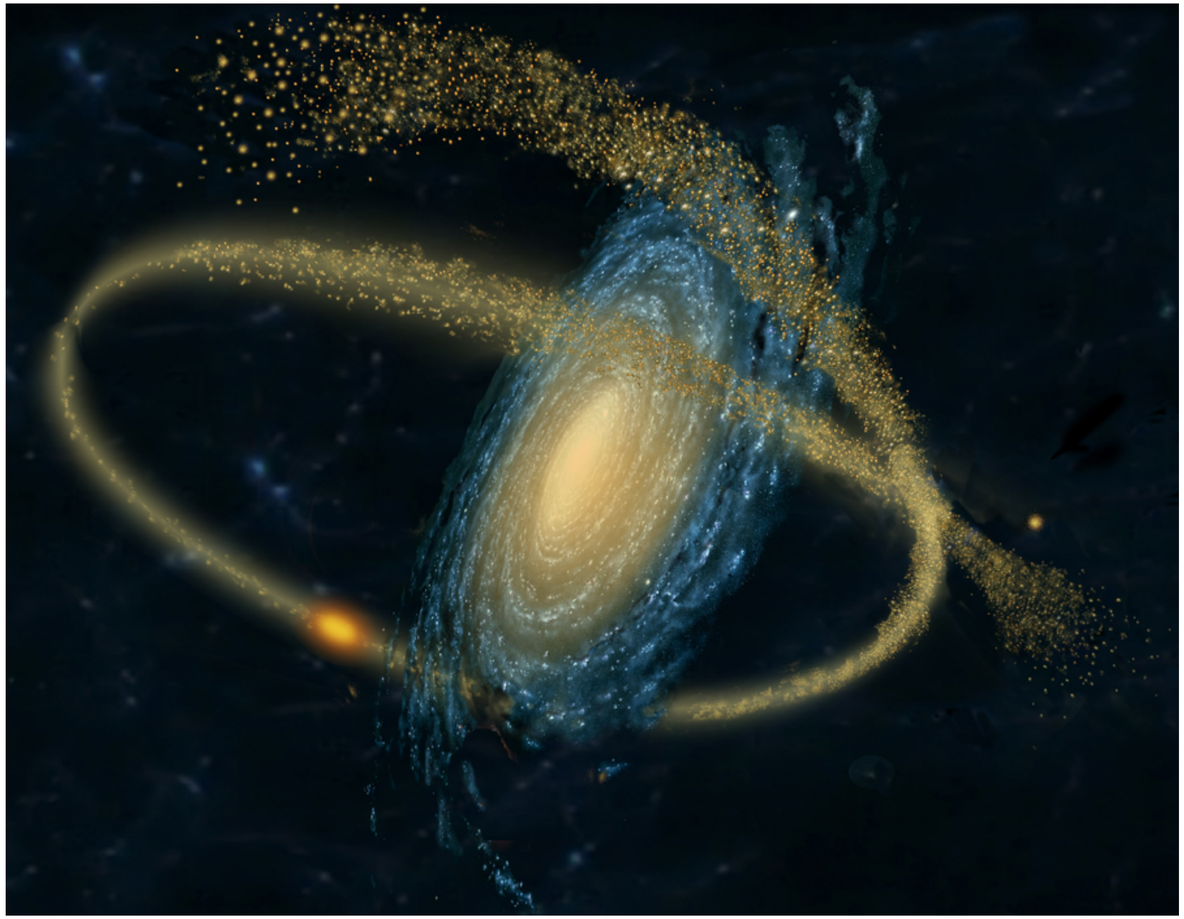


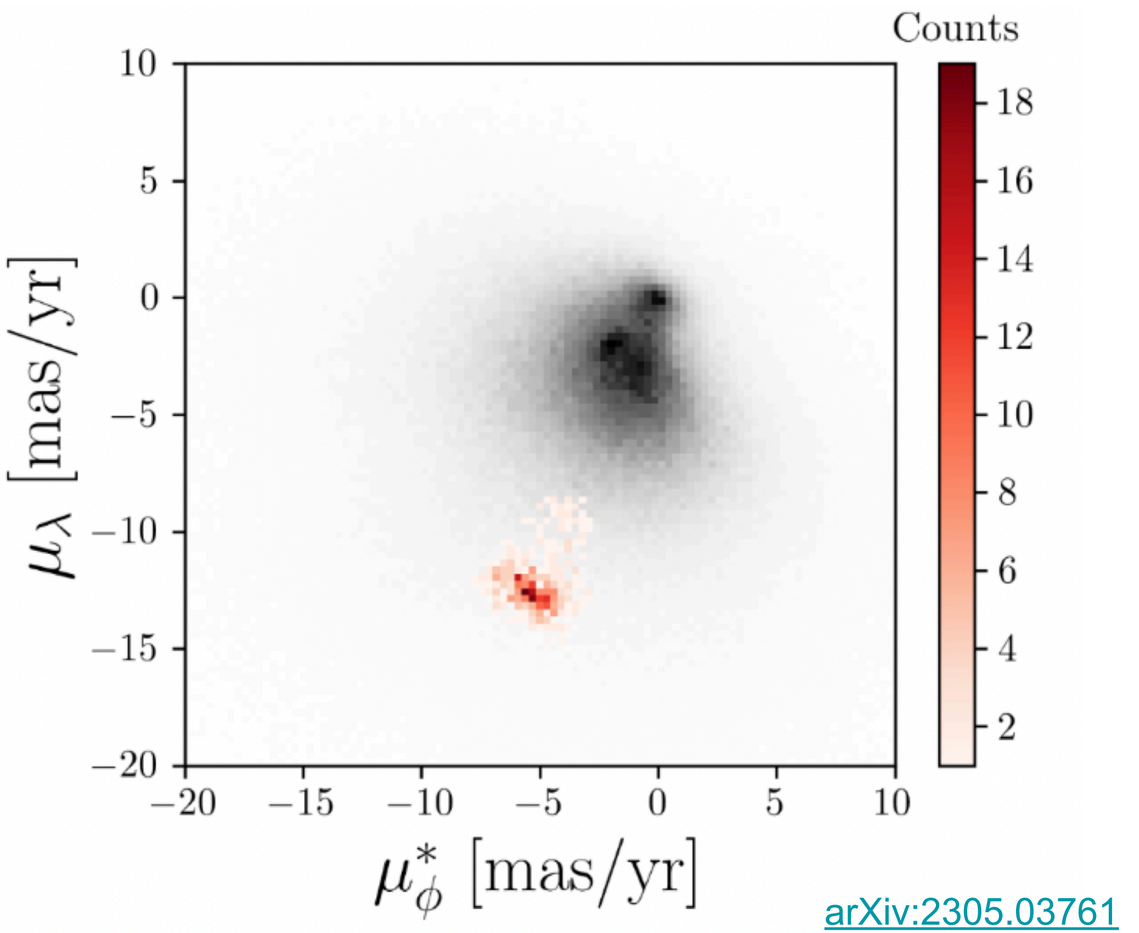
arXiv:2502.14036

arXiv:2508.13566

Anomaly detection in particle physics targets group anomalies, not single-event outliers.

(and group anomalies show up in astrophysics, too!)

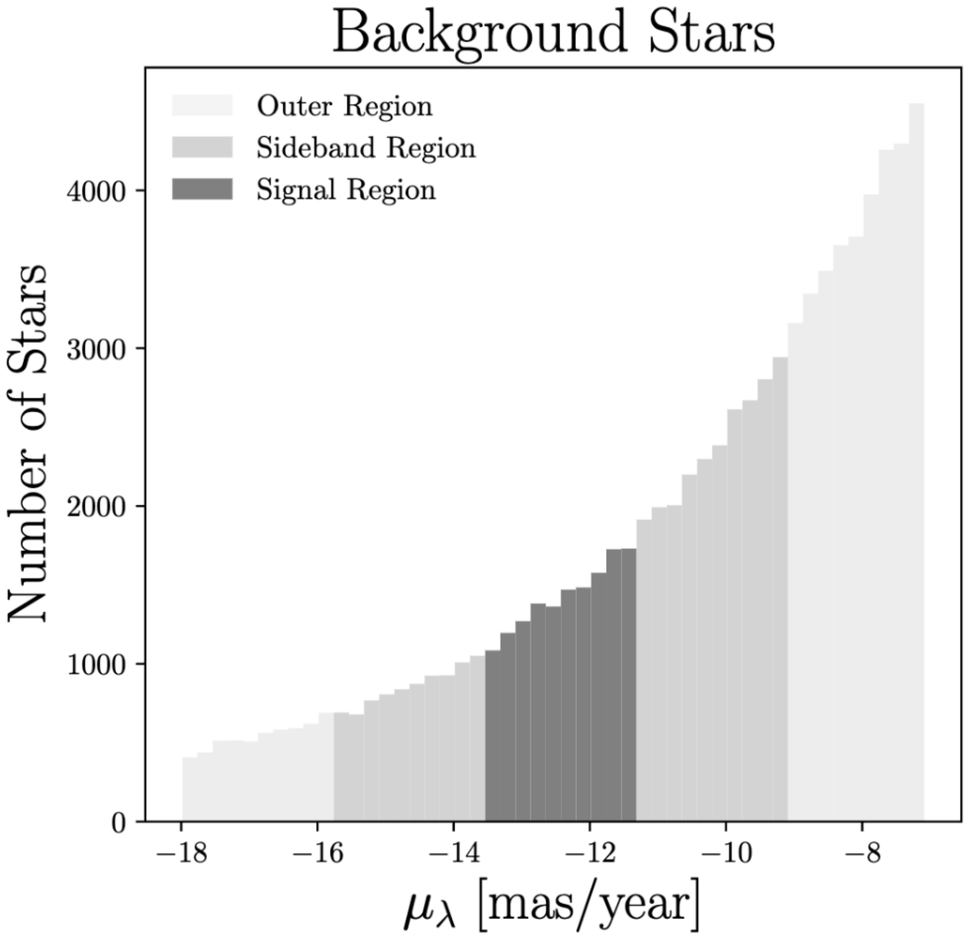


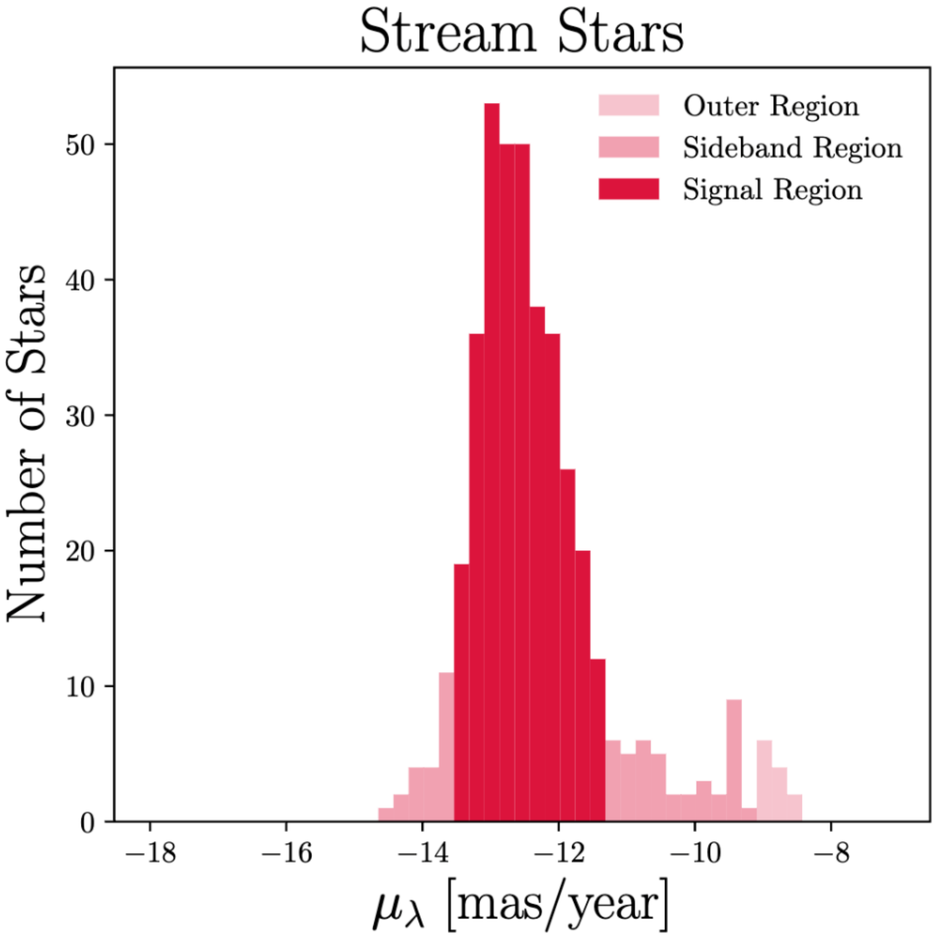


arXiv:2305.03761

Anomaly detection in particle physics targets group anomalies, not single-event outliers.

(and group anomalies show up in astrophysics, too!)

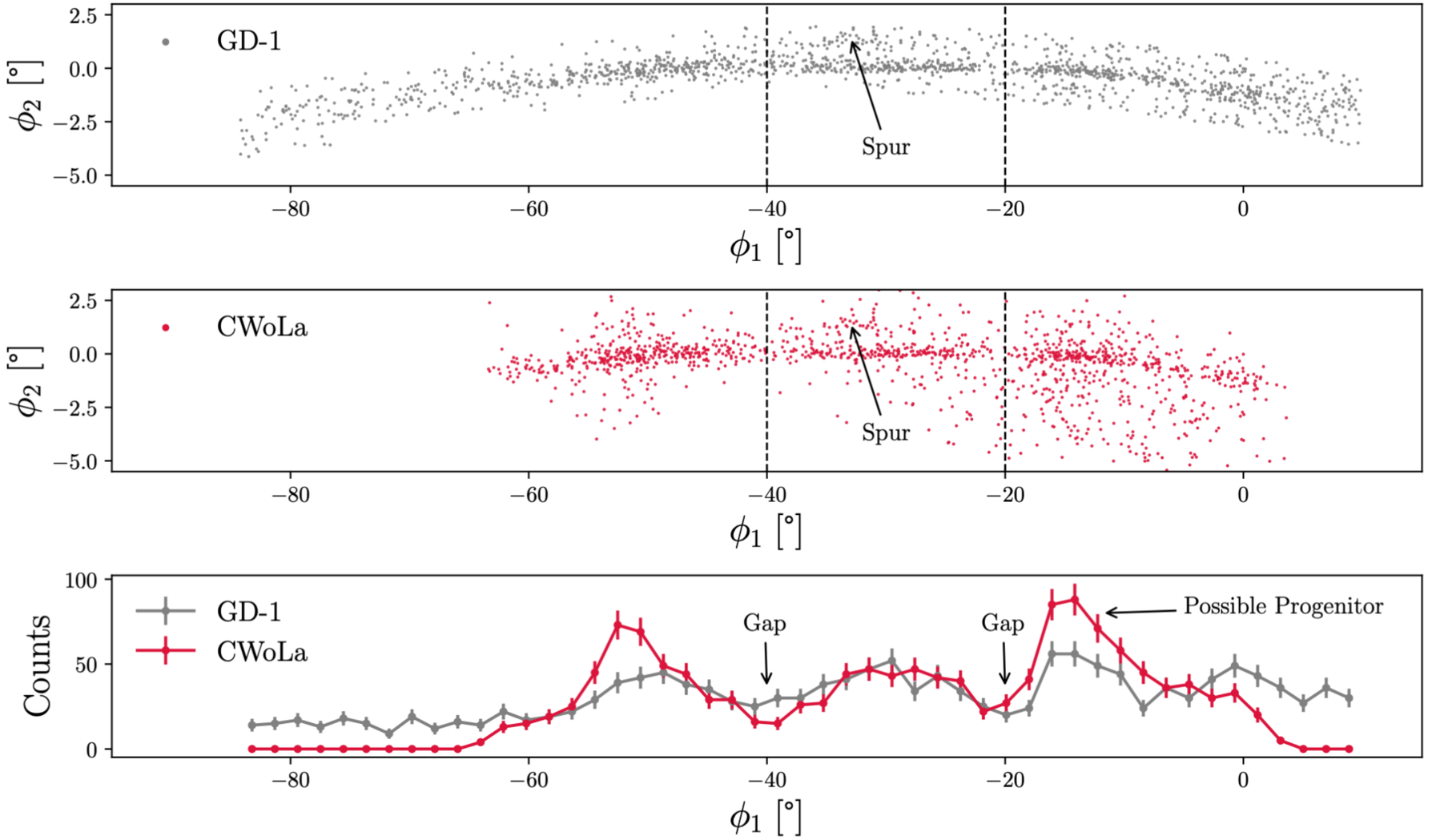




arXiv:2305.03761

Anomaly detection in particle physics targets group anomalies, not single-event outliers.

(and group anomalies show up in astrophysics, too!)



arXiv:2305.03761

Physics data has some qualities that are relatively under-explored in mainstream ML.

Deep theoretical grounding → exploit known symmetries & simulations

Diverse data formats → custom vs. generic embeddings

Detector-specific biases → publish data without detector effects

Group anomalies → statistical anomaly detection

Highly multi-modal → unify inputs across an entire detector

Informed by universal physical laws → multi-detector/disciplinary analyses

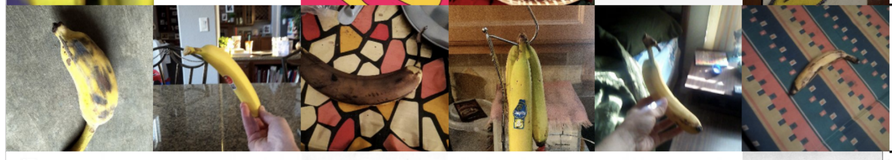
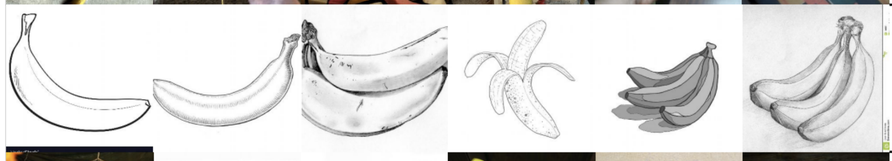
A **modality** is a specific way of perceiving the world – it’s a concept deeply rooted in our own five senses.

(In ML, common modalities include text, audio, video, etc.)

AI industry has multimodal models, but the modalities are relatively few and well-defined:

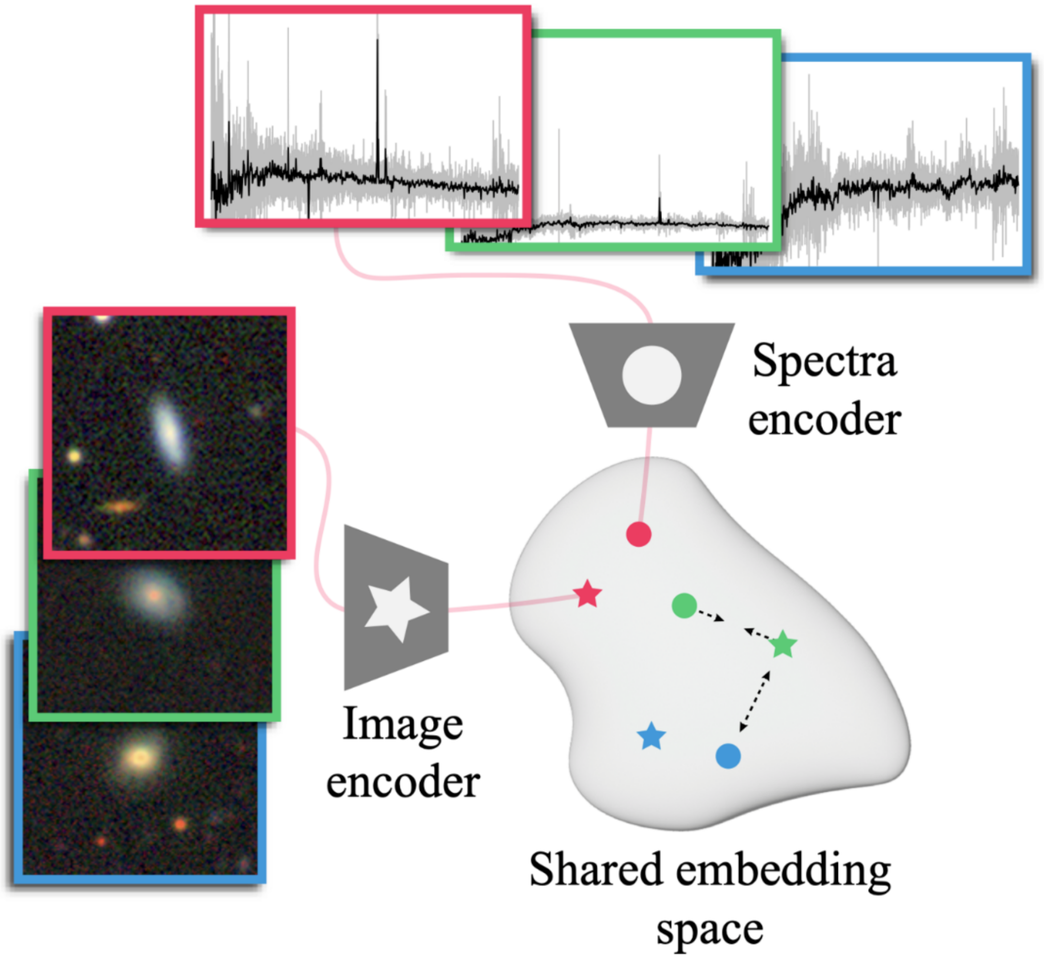
Multimodal models see promising performance in industry and in science.

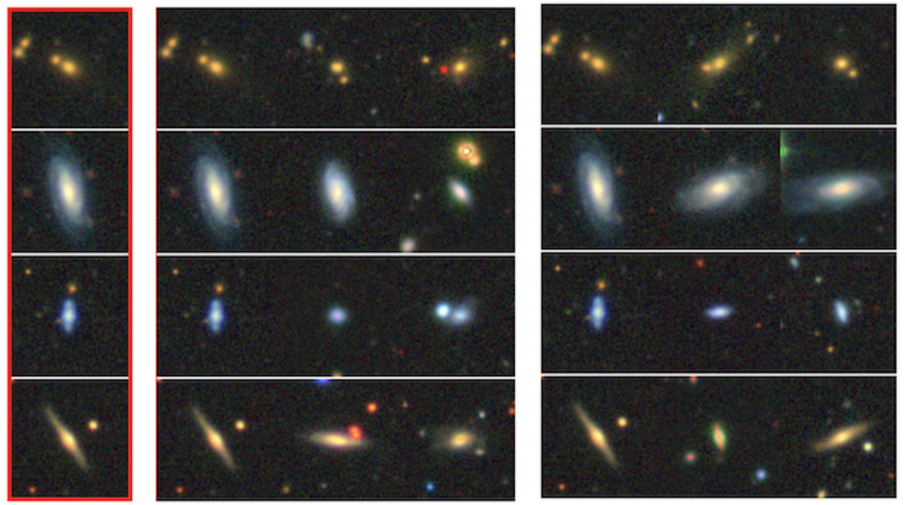
Significant evidence suggests that including multiple views/ modalities in a single model can improve performance on downstream tasks relative to unimodal models:

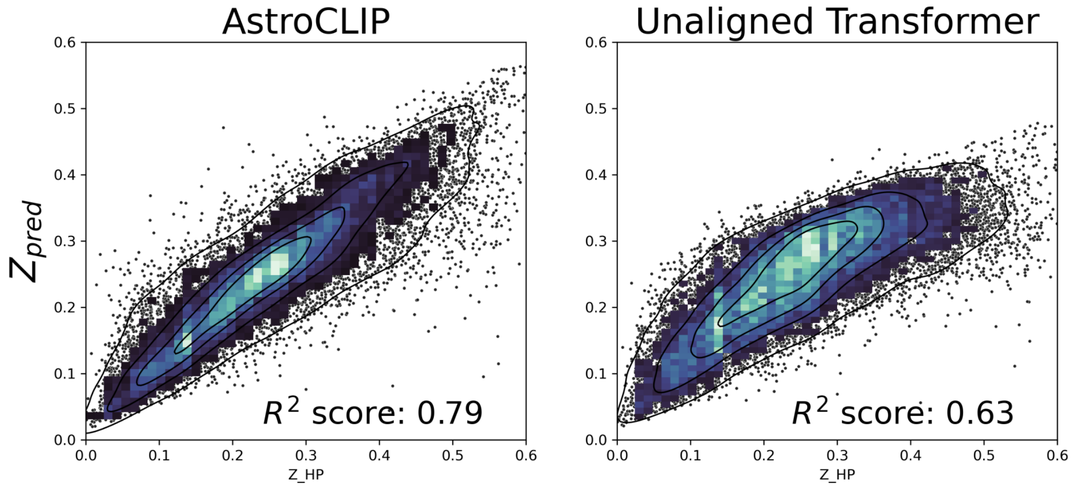
	Dataset Examples	ImageNet ResNet101	Zero-Shot CLIP	Δ Score
ImageNet		**76.2**	**76.2**	0%
ImageNetV2		64.3	**70.1**	+5.8%
ImageNet-R		37.7	**88.9**	+51.2%
ObjectNet		32.6	**72.3**	+39.7%
ImageNet Sketch		25.2	**60.2**	+35.0%
ImageNet-A		2.7	**77.1**	+74.4%

(CLIP, 2021)

Multimodal models see promising performance in industry and in science.

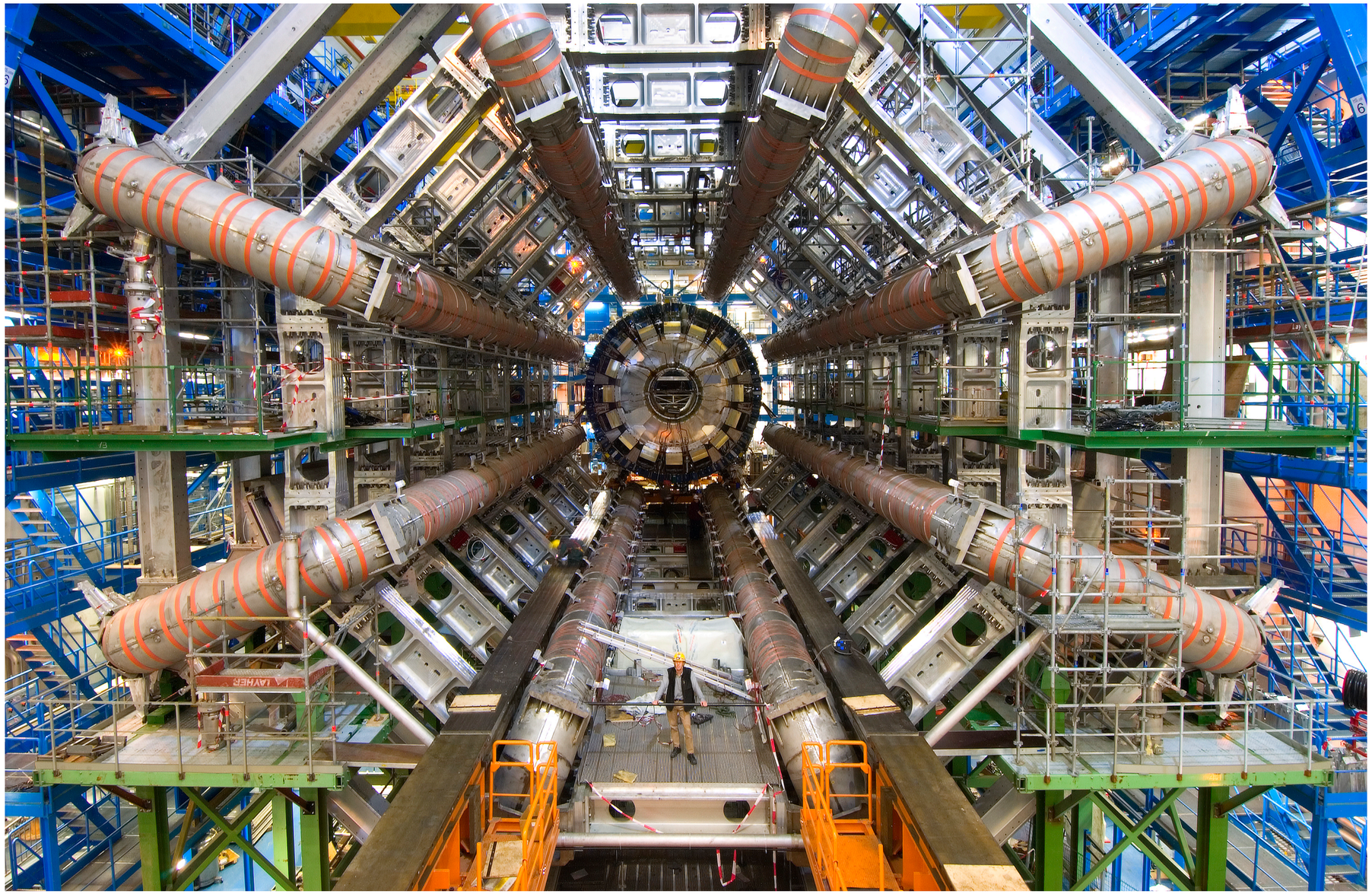






(AstroCLIP, 2023)

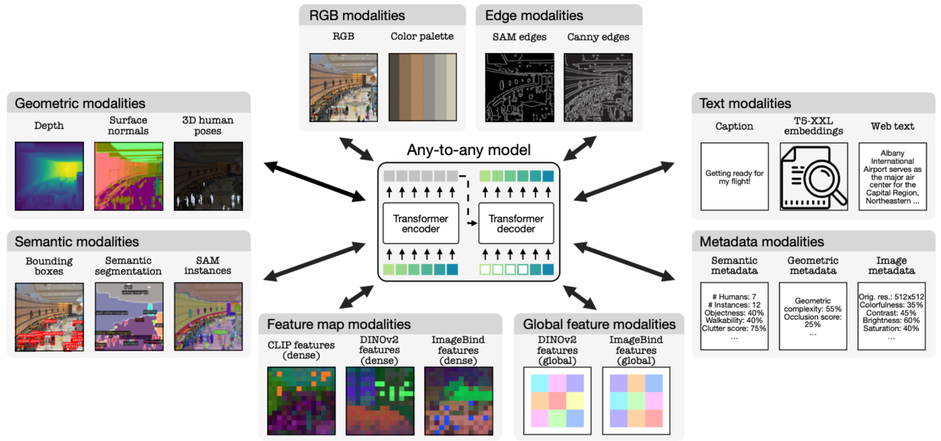
But what happens when we try to encode 10, 20, 50+ modalities at once?



But what happens when we try to encode 10, 20, 50+ modalities at once?

4M-21: An Any-to-Any Vision Model for Tens of Tasks and Modalities

Roman Bachmann[1†*] **Oğuzhan Fatih Kar**[1*] **David Mizrahi**[2†*] **Ali Garjani**[1]
Mingfei Gao[2] **David Griffiths**[2] **Jiaming Hu**[2] **Afshin Dehghan**[2] **Amir Zamir**[1]
[1]Swiss Federal Institute of Technology Lausanne (EPFL) [2]Apple
https://4m.epfl.ch

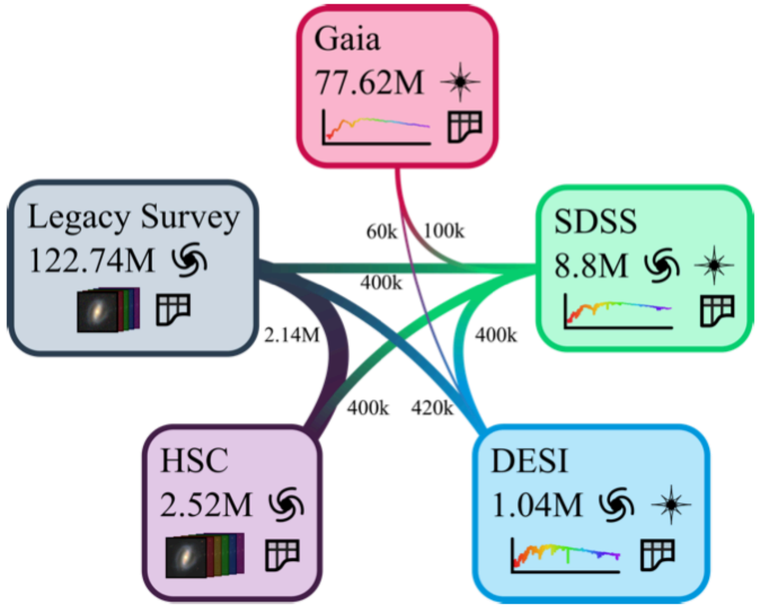


arXiv:2406.09406

Polymathic

AION-1: Omnimodal Foundation Model for Astronomical Sciences

Liam Parker[*,1,2,3,4], **Francois Lanusse**[*,5,2], **Jeff Shen**[*,6], **Ollie Liu**[7], **Tom Hehir**[8], **Leopoldo Sarra**[3], **Lucas Meyer**[3], **Micah Bowles**[9], **Sebastian Wagner-Carena**[2,3], **Helen Qu**[2], **Siavash Golkar**[2,3], **Alberto Bietti**[2], **Hatim Bourfoune**[10], **Nathan Cassereau**[10], **Pierre Cornette**[10], **Keiya Hirashima**[2,11], **Geraud Krawezik**[2], **Ruben Ohana**[2], **Nicholas Lourie**[3], **Michael McCabe**[2,3], **Rudy Morel**[2], **Payel Mukhopadhyay**[1,8], **Mariel Pettee**[12], **Bruno Regaldo-Saint Blancard**[2], **Kyunghyun Cho**[3], **Miles Cranmer**[8], **Shirley Ho**[2,3,6]



arXiv:2510.17960

But what happens when we try to encode 10, 20, 50+ modalities at once?

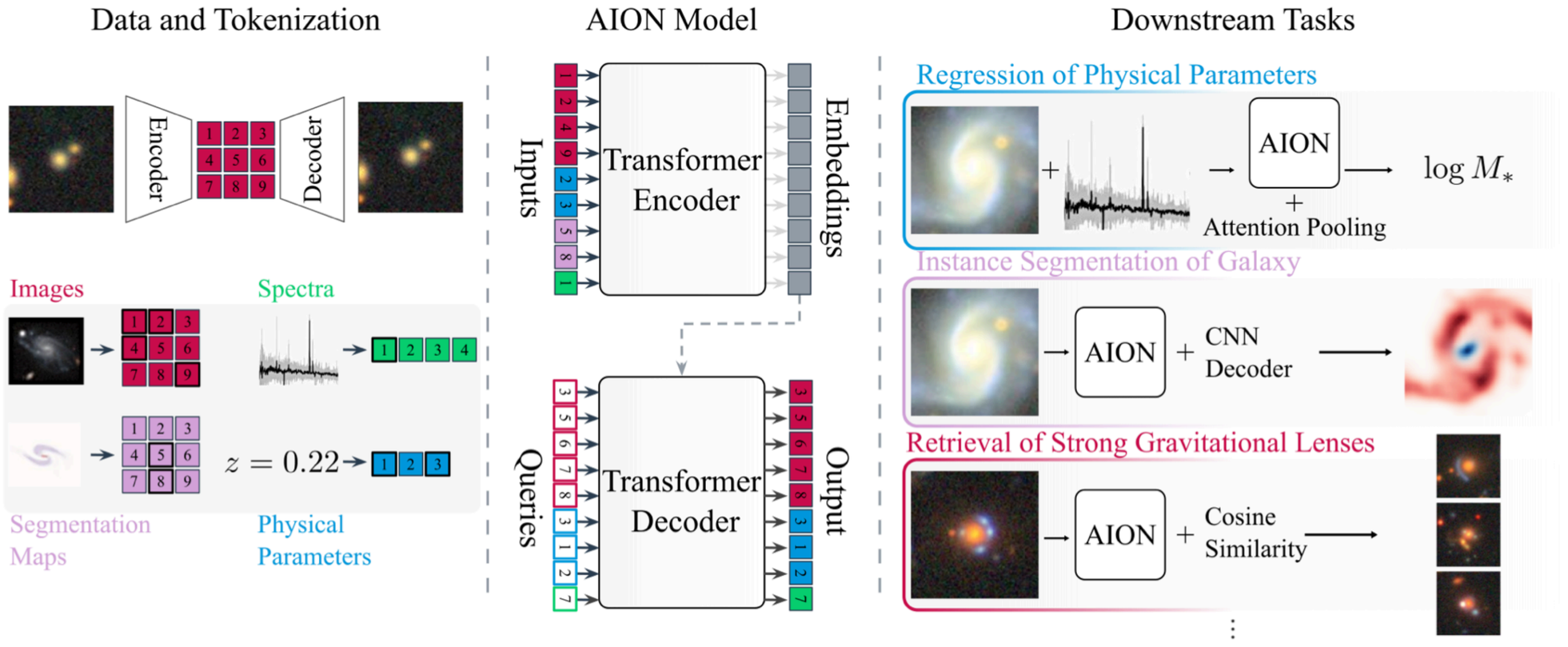


Figure 1: AION-1 integrates 39 different data modalities — multiband images, optical spectra, and various properties and measurements — into a single model usable for a wide range of downstream applications. It implements a two-step process: first, bespoke tokenization strategies that homogenize the diverse scientific data, followed by multimodal masked modeling that learns how different observations relate, inducing a deep understanding of the underlying physical objects. Astronomers can then leverage AION-1's rich astrophysical understanding for a variety of downstream tasks.

arXiv:2510.17960

Physics data has some qualities that are relatively under-explored in mainstream ML.

Deep theoretical grounding → exploit known symmetries & simulations

Diverse data formats → custom vs. generic embeddings

Detector-specific biases → publish data without detector effects

Group anomalies → statistical anomaly detection

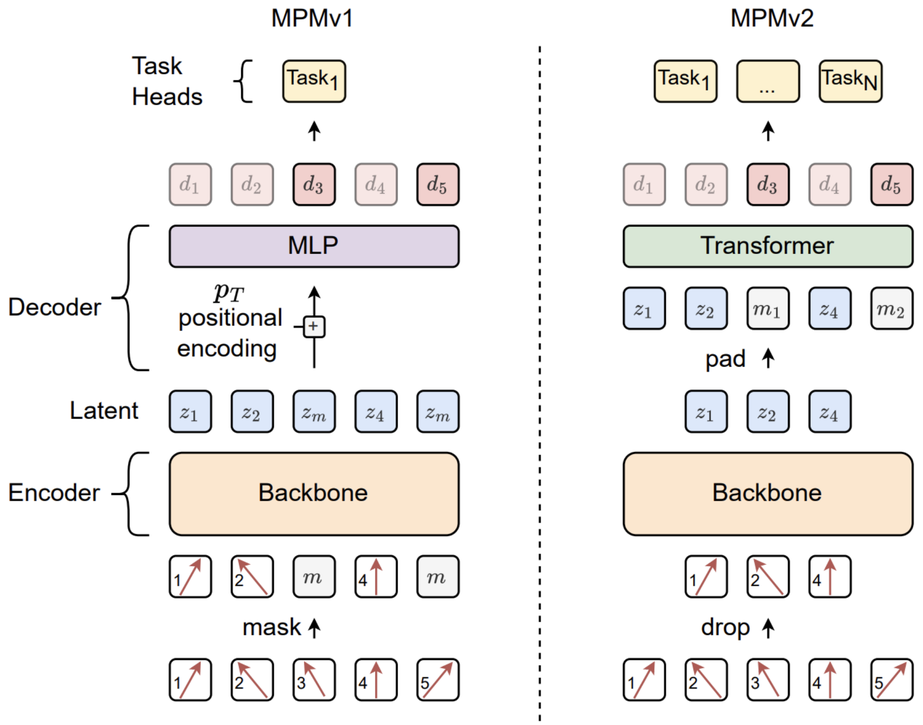
Highly multi-modal → unify inputs across an entire detector

Informed by universal physical laws → multi-detector/disciplinary analyses

We've started to see some domain-specific foundation models in physics...

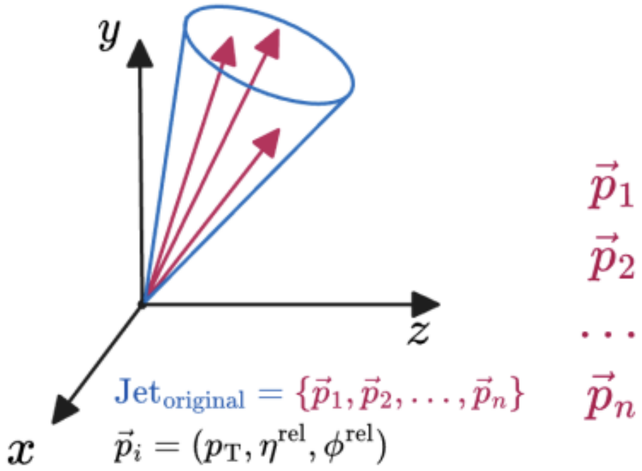
Masked Particle Modeling (MPM)

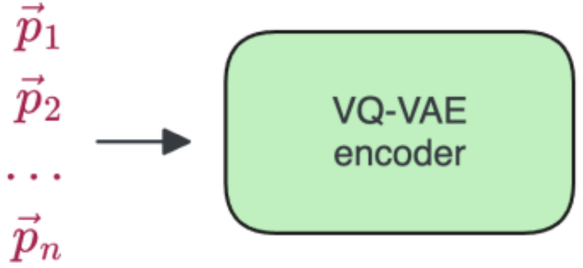
arXiv:2401.13537 & arXiv:2409.12589



Omni-Jet α

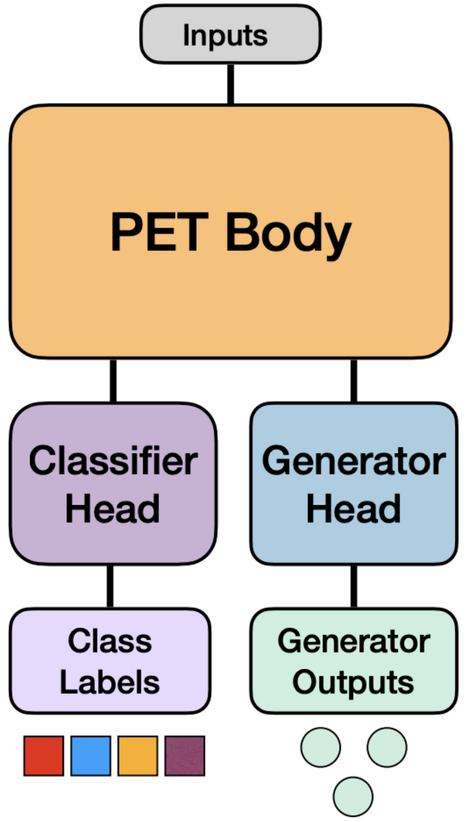
arXiv:2403.05618





OmniLearn

arXiv:2404.16091
arXiv:2502.14652



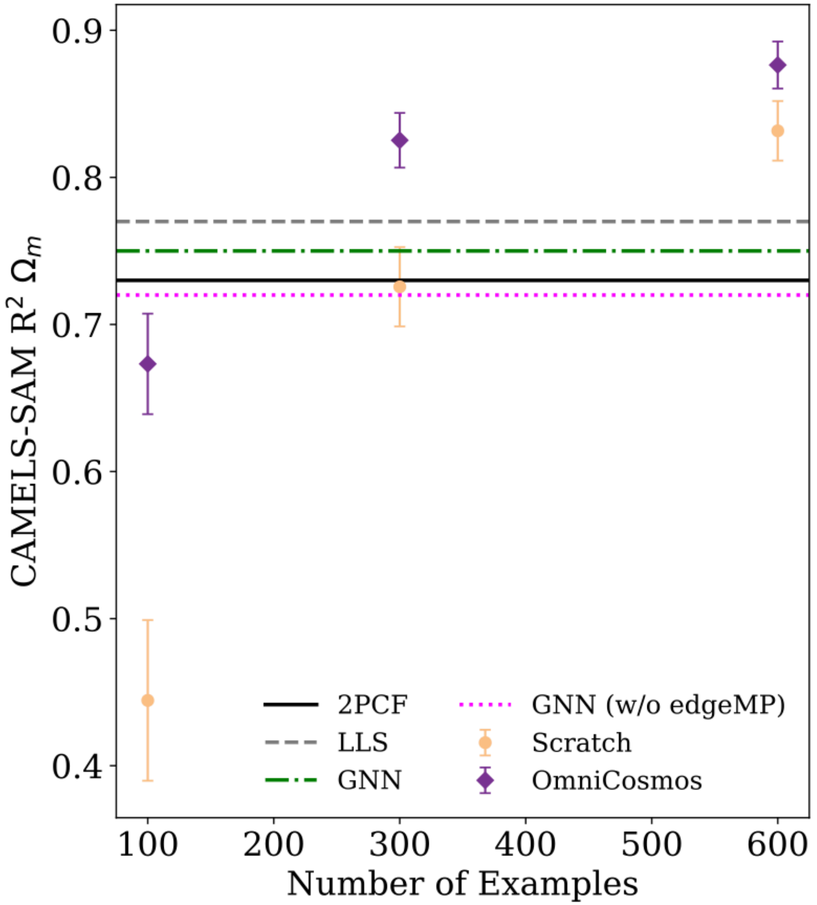
...and it seems like they can be fine-tuned to transfer surprisingly well out-of-domain:

OMNICOSMOS: Transferring Particle Physics Knowledge Across the Cosmos

Vinicius Mikuni,[1,*] Ibrahim Elsharkawy,[2,3,†] and Benjamin Nachman[4,5,‡]

[1] *Nagoya University, Kobayashi-Maskawa Institute, Aichi 464-8602, Japan*
[2] *National Energy Research Scientific Computing Center (NERSC), Lawrence Berkeley National Laboratory, Berkeley, CA, USA*
[3] *Department of Physics, University of Toronto, Toronto, ON, Canada*
[4] *Department of Particle Physics and Astrophysics, Stanford University, Stanford, CA 94305, USA*
[5] *Fundamental Physics Directorate, SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA*

Foundation models build an effective representations of data that can be deployed on diverse downstream tasks. Previous research developed the OMNILEARNED foundation model for collider physics and showed that it could significantly advance discovery potential across collider experiments. In this paper we go beyond collider physics and show that Foundation Models trained on collider data can help improve the prediction of cosmological parameters and to predict halo and galaxy velocities in different datasets from COSMOBENCH. This is the first time a collider physics model is shown to generalize across scientific fields.



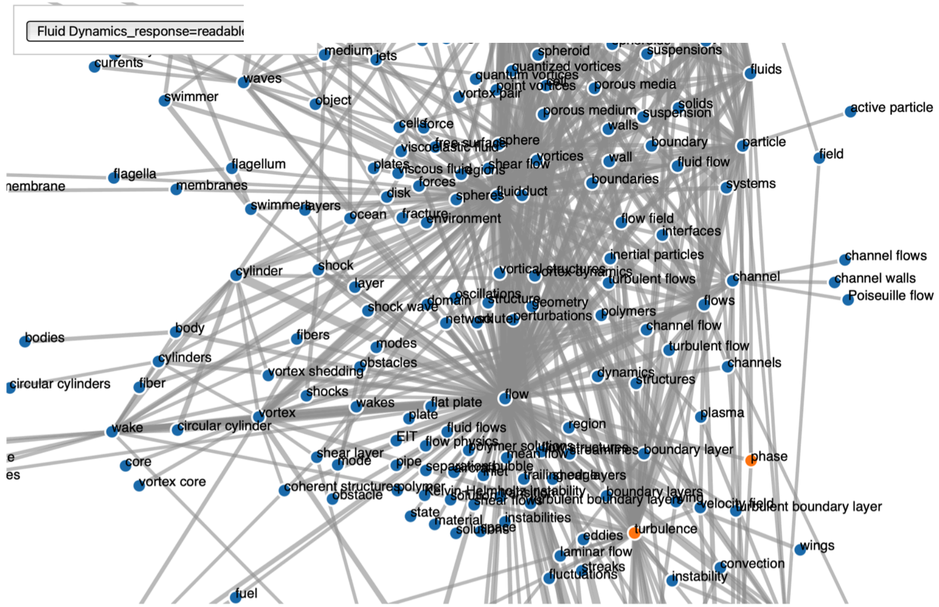
arXiv:2512.24422

Can we better characterize the scientific landscape to redefine “near” and “far”?

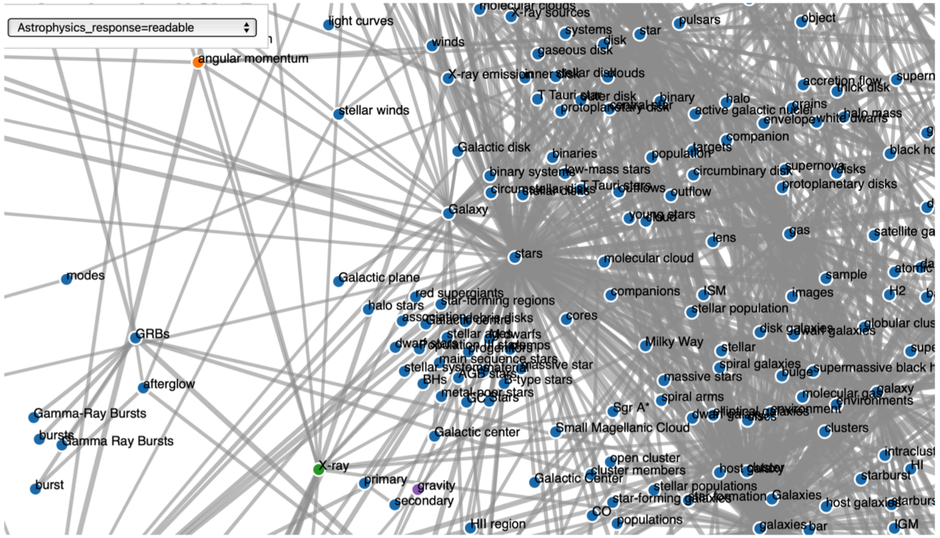
What’s In Your Field?
Mapping Scientific Research with Knowledge Graphs and LLMs

Abhipsha Das[1,2] Nicholas Lourie[1,3] Siavash Golkar[1,3] Mariel Pettee[1,4]

[1]Polymathic AI [2]Flatiron Institute [3]New York University [4]Lawrence Berkeley National Laboratory
abhipsha.das@nyu.edu



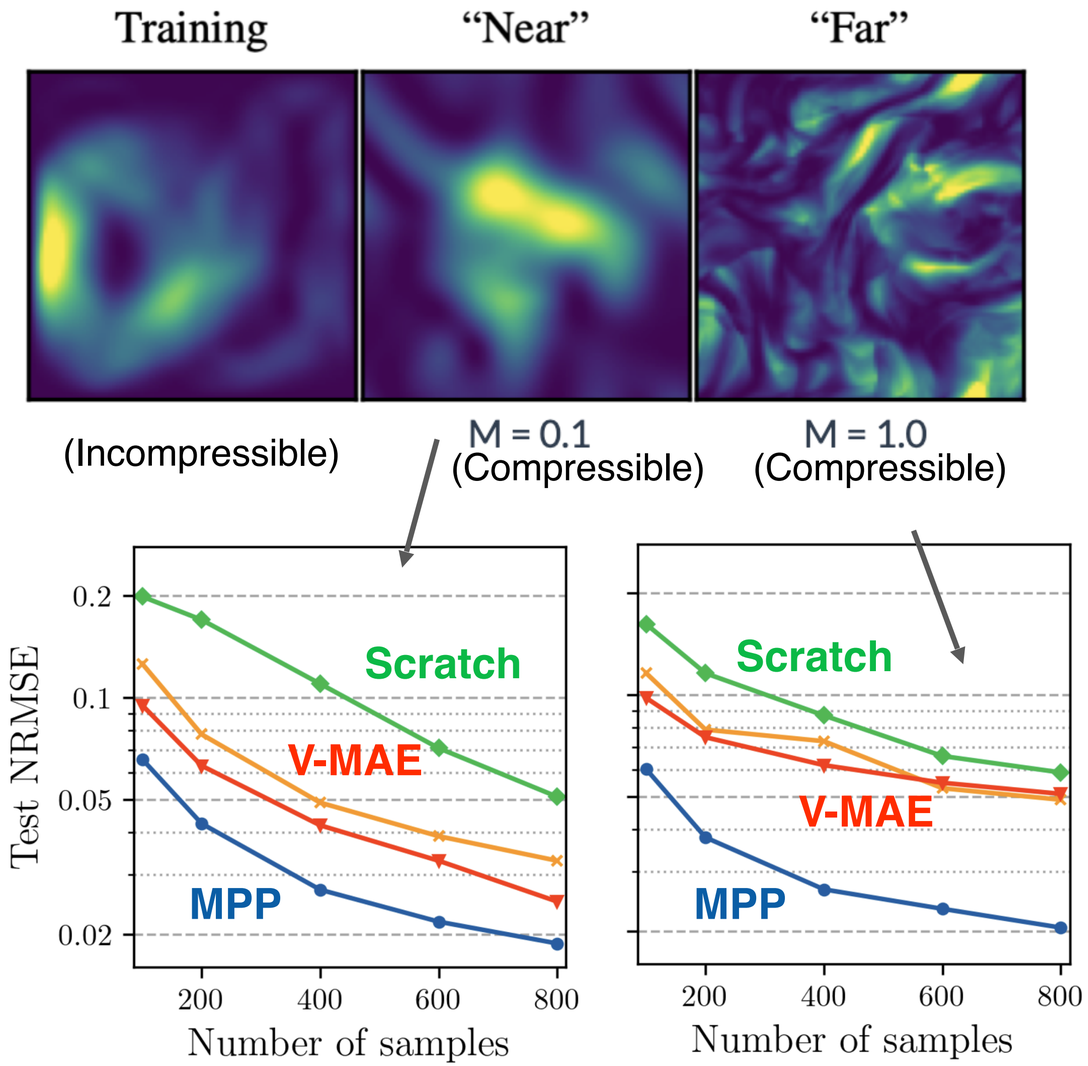
(c) Fluid dynamics clusters centered around "flow," branching into related flow and turbulence-based physical phenomena.



(d) Astrophysical objects and phenomena spanning multiple scales, from individual stars and binaries to galactic structures and clusters.

arxiv:2503.09894

Will our analyses benefit from incorporating information from other datasets?



arXiv:2310.02994

The Multimodal Universe: Enabling Large-Scale Machine Learning with 100 TB of Astronomical Scientific Data

MULTIMODAL UNIVERSE

arXiv:2412.02527

How much does the shared information content affect downstream performance?

Multimodal Datasets with Controllable Mutual Information

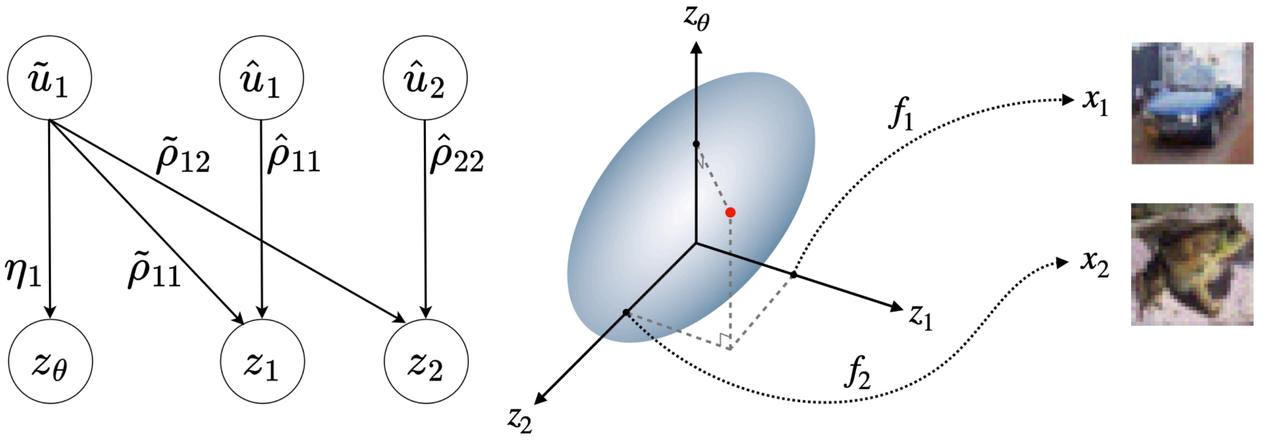
Raheem Karim Hashmani
University of Wisconsin–Madison
hashmani@wisc.edu

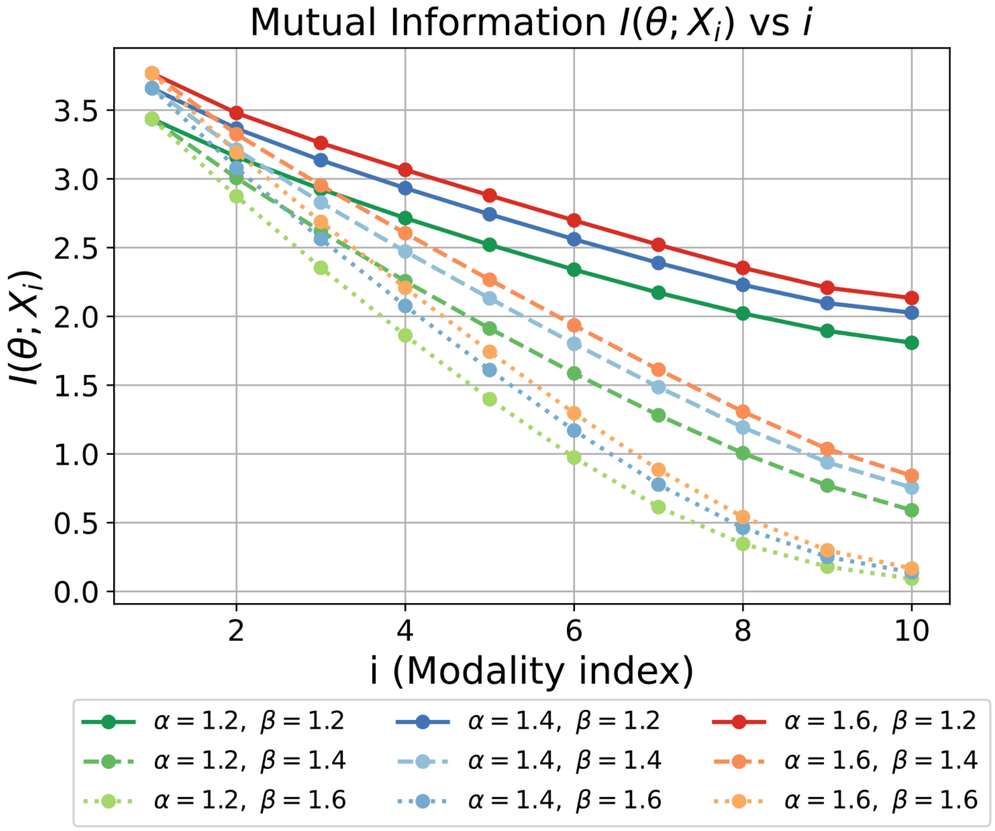
Garrett W. Merz
University of Wisconsin–Madison
garrett.merz@wisc.edu

Helen Qu
Flatiron Institute
hqu@flatironinstitute.org

Mariel Pettee
University of Wisconsin–Madison
mpettee@wisc.edu

Kyle Cranmer
University of Wisconsin–Madison

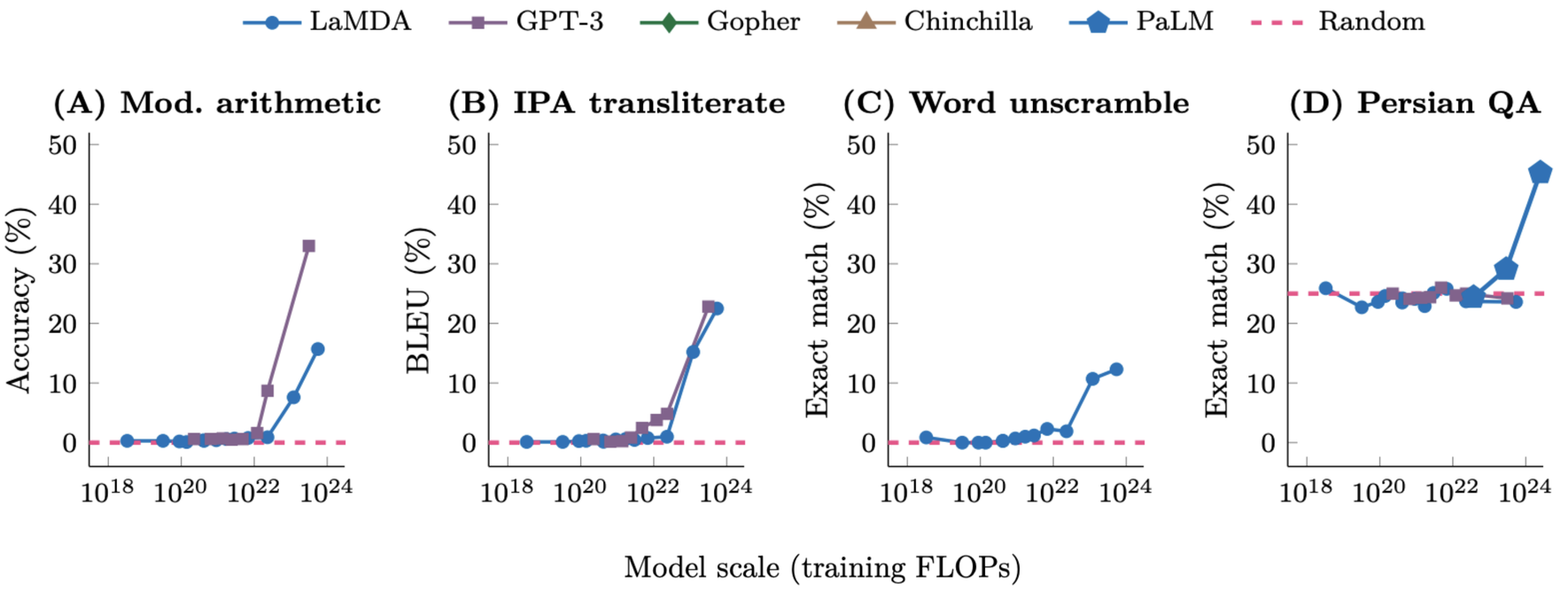




arXiv:2406.09406

Will our analyses benefit from incorporating information from other datasets?

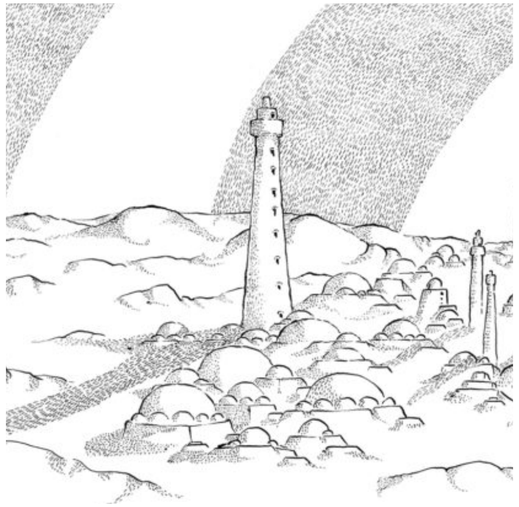
ChatGPT and other LLMs were trained on O(1%) of the Internet. Their powerful generalization capabilities emerged at some point due to a combination of model capacity, heterogeneity, and training dataset size.

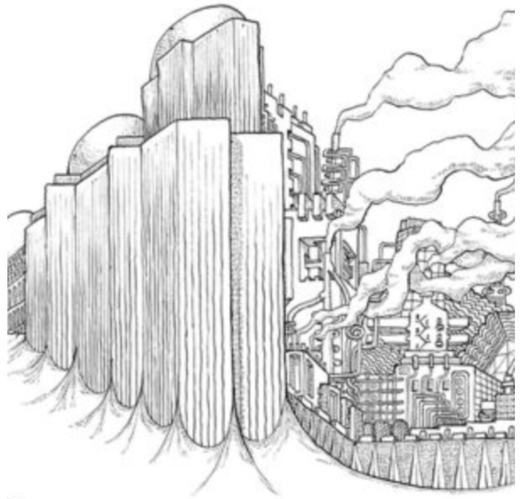


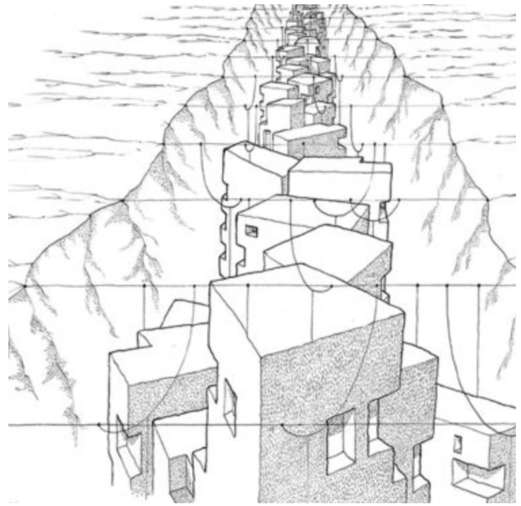
arXiv:2206.07682

Will scientific foundation models trained on vast, heterogeneous datasets also yield emergent out-of-the-box generalization properties, as we see in many industry-scale foundation models?

"Every time I describe a city I am saying something about Venice."

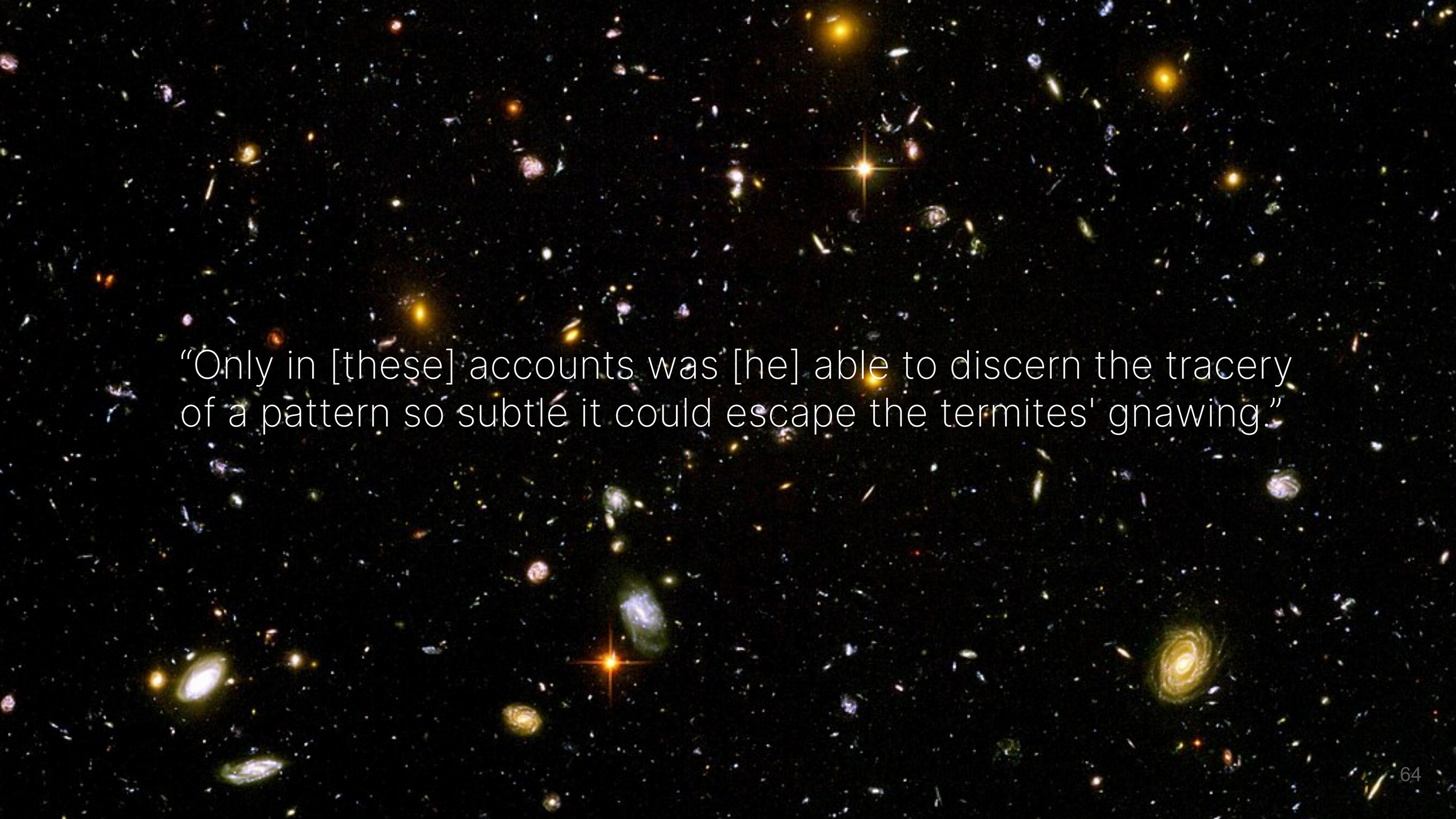


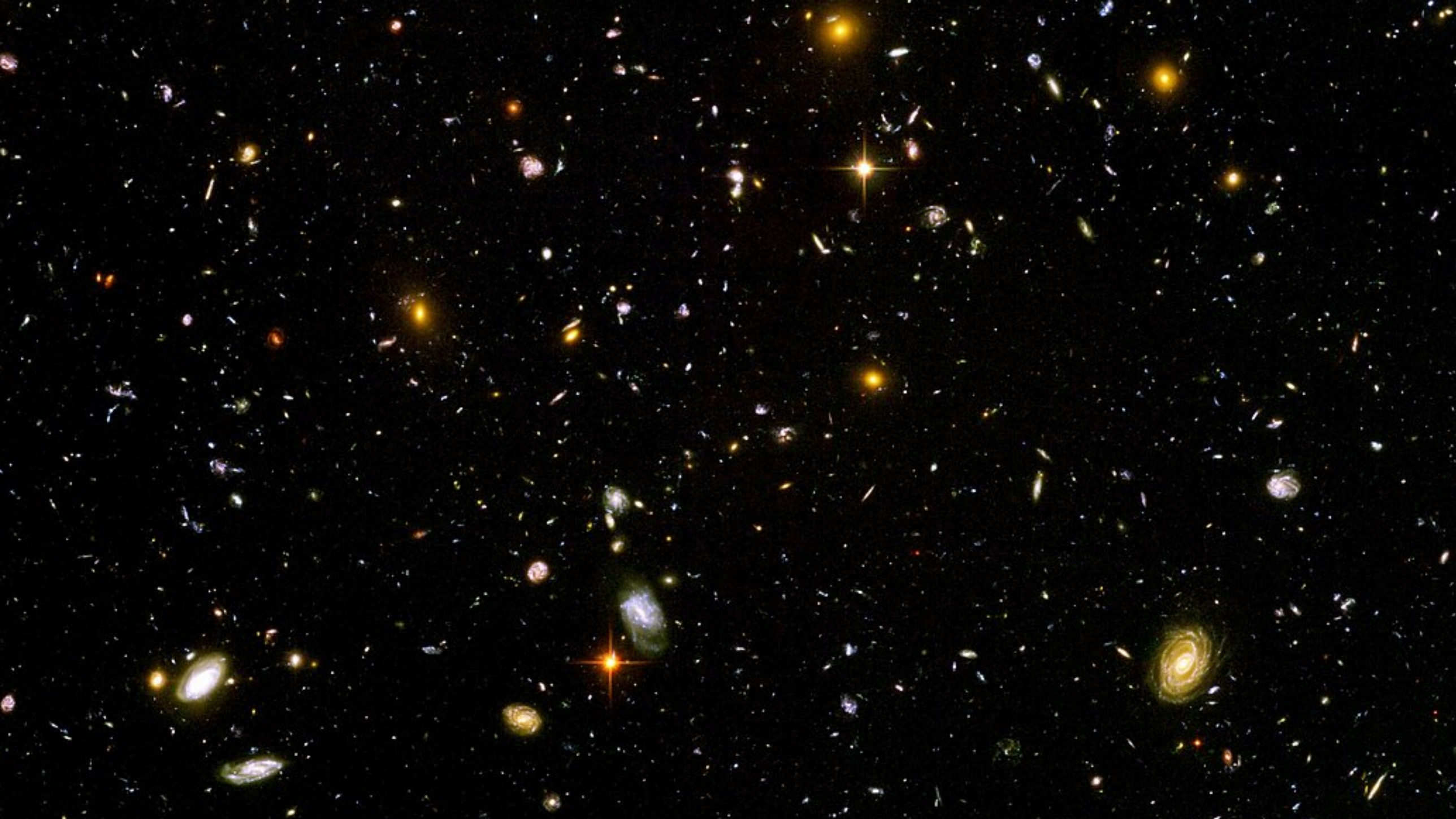




Art by Matteo Pericoli

“Only in [these] accounts was [he] able to discern the tracery of a pattern so subtle it could escape the termites' gnawing.”


“Only in [these] accounts was [he] able to discern the tracery
of a pattern so subtle it could escape the termites' gnawing.”



Classifier functions can be re-used to directly approximate a likelihood ratio.

A vanilla NN classifying between two classes could be trained using **binary cross-entropy loss:**

$$\mathcal{L}_{\mathrm{BCE}}[f] = -\int d\vec{x}\,(p_A(\vec{x})\log(f(\vec{x})) + p_B(\vec{x})\log(1-f(\vec{x})))$$

The function that minimizes this expression will satisfy, for any small variatic $\delta f(\vec{x})$:

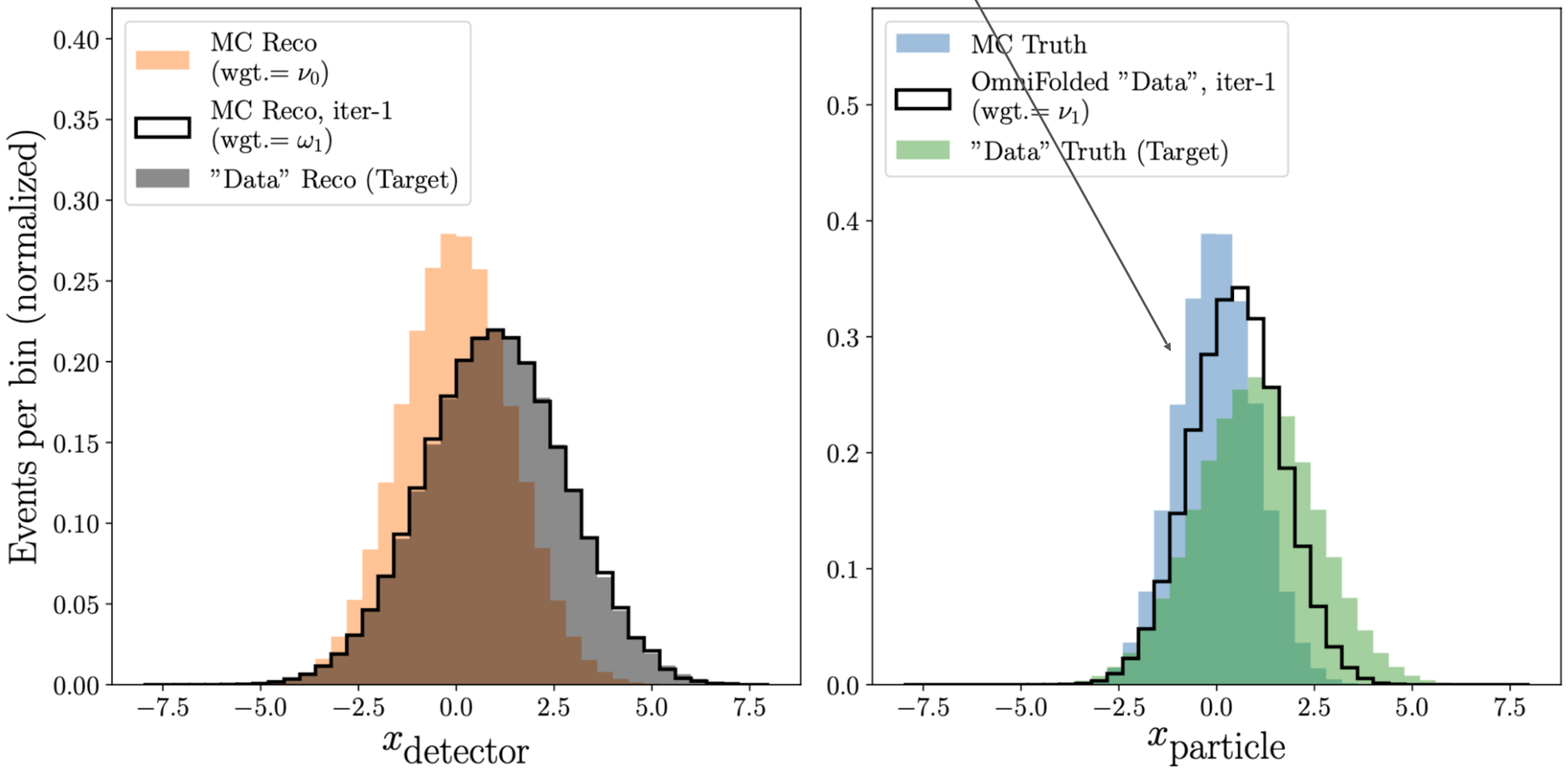
$$\delta\mathcal{L}_{\mathrm{BCE}} = \int d\vec{x}\left(-\frac{p_A(\vec{x})}{f(\vec{x})} + \frac{p_B(\vec{x})}{1-f(\vec{x})}\right)\delta f(\vec{x}) = 0$$

$$-\frac{p_A(\vec{x})}{f(\vec{x})} + \frac{p_B(\vec{x})}{1-f(\vec{x})} = 0$$

Rescaling of classifier outpu $\quad \frac{f(\vec{x})}{1-f(\vec{x})} = \frac{p_A(\vec{x})}{p_B(\vec{x})}$ Likelihood ratio

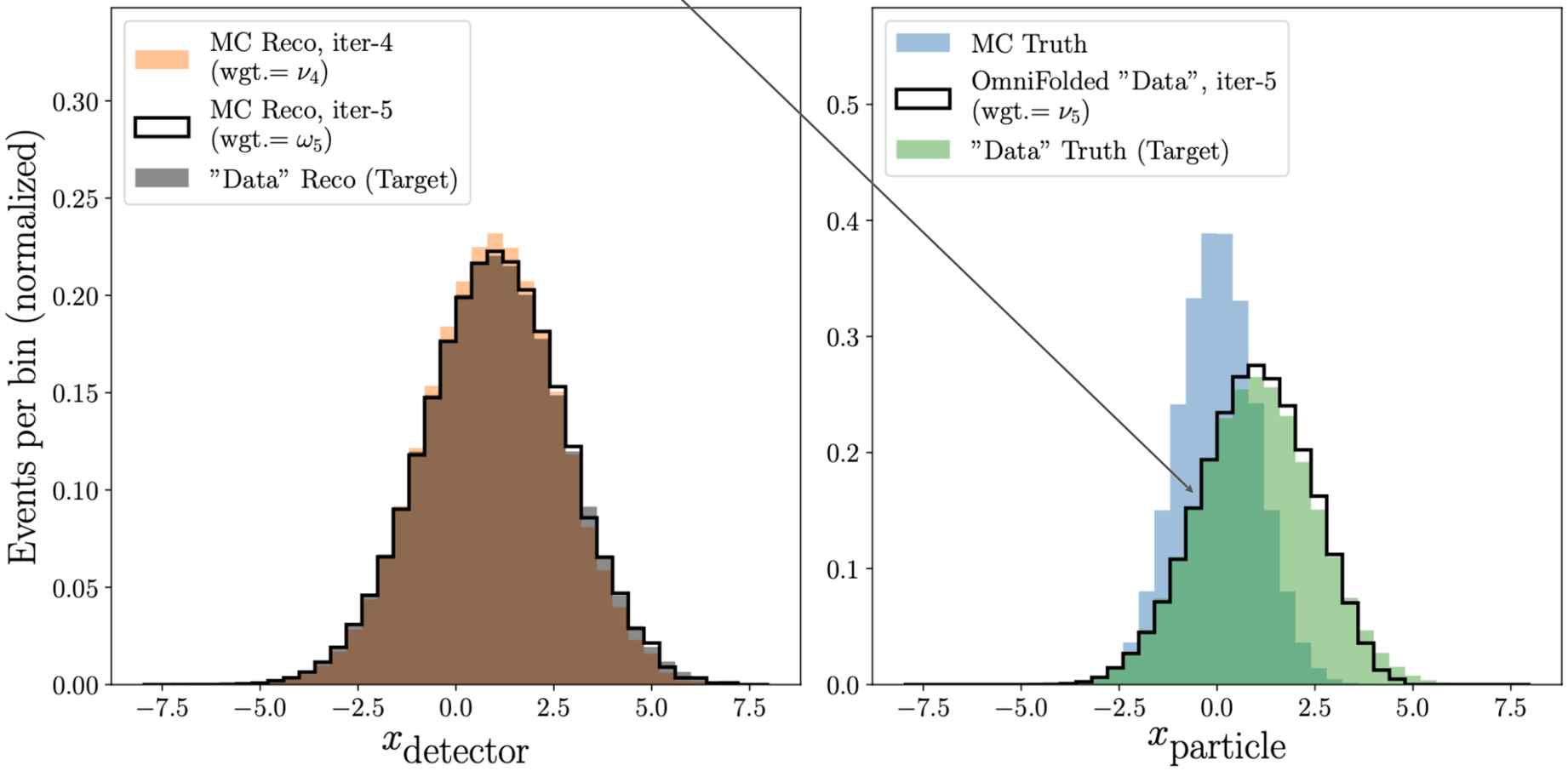
OmniFold reweights truth-level MC to estimate “truth-level” data.

After one iteration of our method, the **truth-level MC** is **reweighted** to look more like **truth-level data**.



OmniFold reweights truth-level MC to estimate “truth-level” data.

After five iterations of our method, the **reweighted** truth-level MC is closely aligned with truth-level data.



The OmniFold procedure requires two datasets* as inputs:

(*In practice, we use samples from different MC generators as well as systematically-shifted samples to determine uncertainties.)

- MC sample with events at both detector-level and particle-level
- Real data

In a multi-stage and iterative process, a series of neural networks are trained to learn a reweighting function that maps particle-level MC distributions to particle-level data distributions.