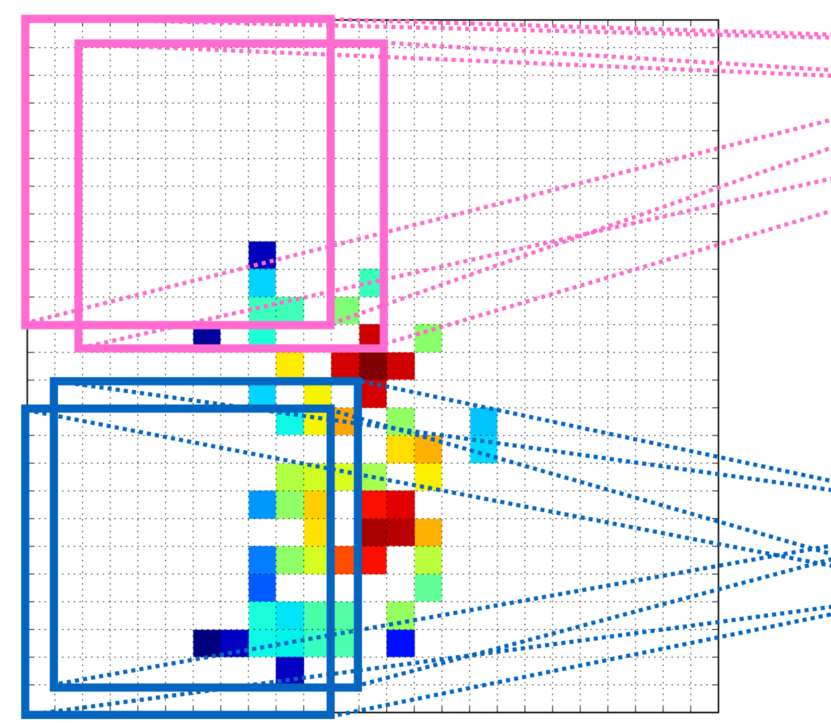
Unfolding:
Machine Learning Approaches

Benjamin Nachman

Lawrence Berkeley National Laboratory

bpnachman.com
bpnachman@lbl.gov

@bpnachman bnachman



AI4EIC
Nov. 2023

Unfolding

Deconvolution (“unfolding”):
correcting for detector effects

Unfolding

Deconvolution ("unfolding"):
correcting for detector effects

Key aspect of all cross section measurements, across particle/nuclear/astro physics (!)

Unfolding

Proton-Proton

Nucleus-Nucleus

Electron-Proton

Neutrino-Nucleus

Cosmic Rays

Electron-Positron

Particle/Nuclear/Astro Physics Experiments

Deconvolution ("unfolding"): correcting for detector effects

Key aspect of **all** cross section measurements, across particle/nuclear/astro physics (!)

Unfolding

Proton-Proton

Nucleus-Nucleus

Electron-Proton

Neutrino-Nucleus

Cosmic Rays

Electron-Positron

Particle/Nuclear/Astro Physics Experiments

Deconvolution ("unfolding"): correcting for detector effects

Key aspect of **all** cross section measurements, across particle/nuclear/astro physics (!)

Why "unfold" instead of "fold"?

Unfolding is ill-posed, BUT only way to compare different experiments and to compare with non fully exclusive predictions. Data also survive much longer.

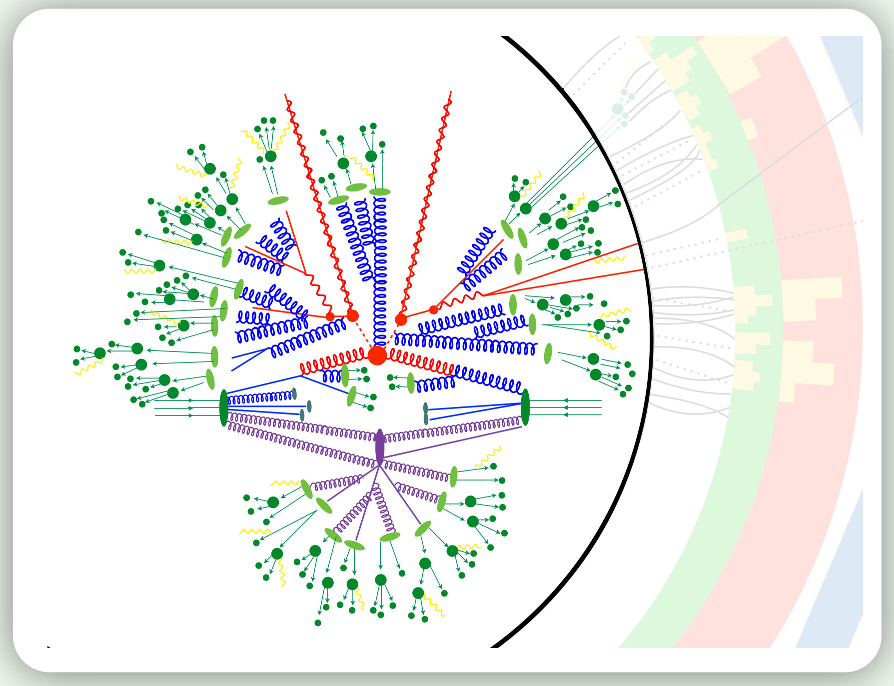
The Unfolding Challenge

2203.16722

The Unfolding Challenge

Particle Level

Want this



2203.16722

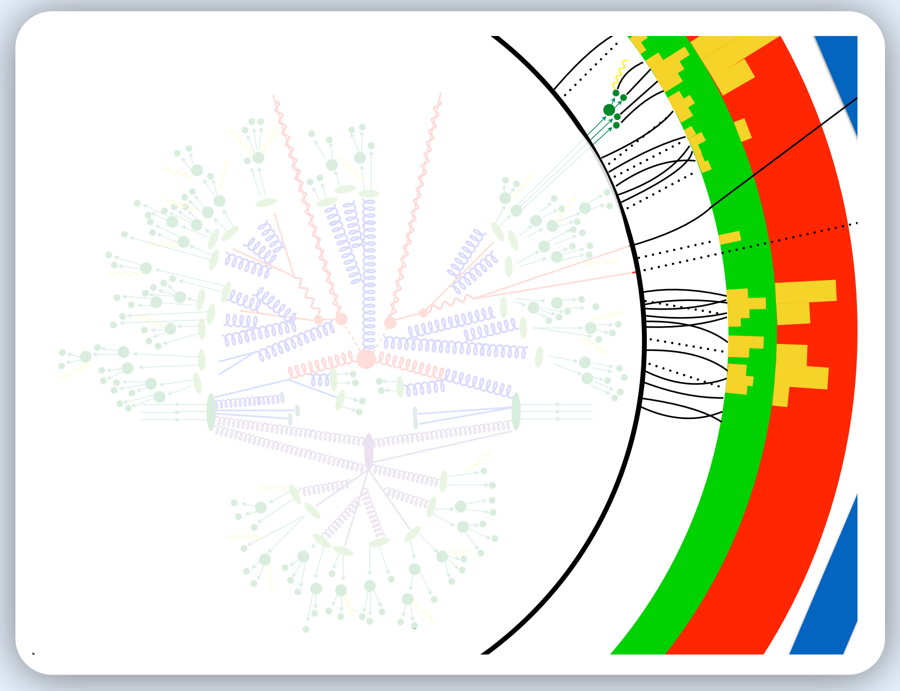
The Unfolding Challenge

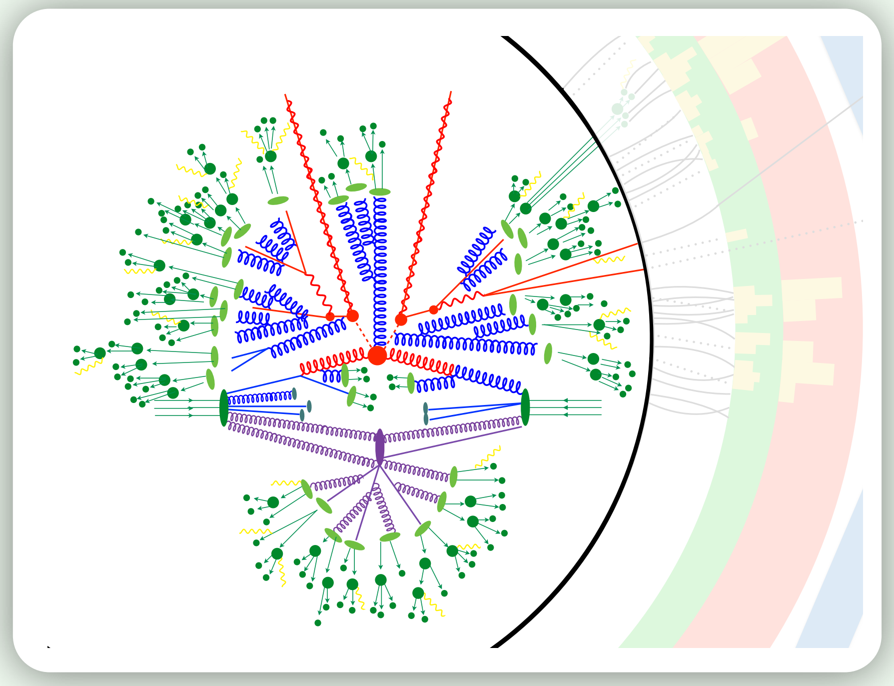
Detector Level

Particle Level

Measure this

Want this





2203.16722

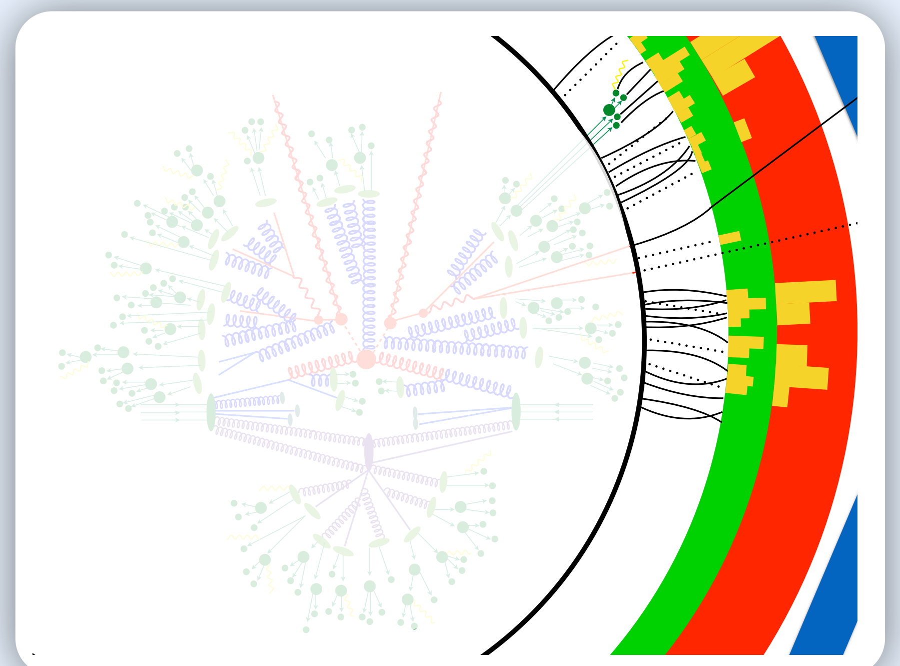
Detector Level

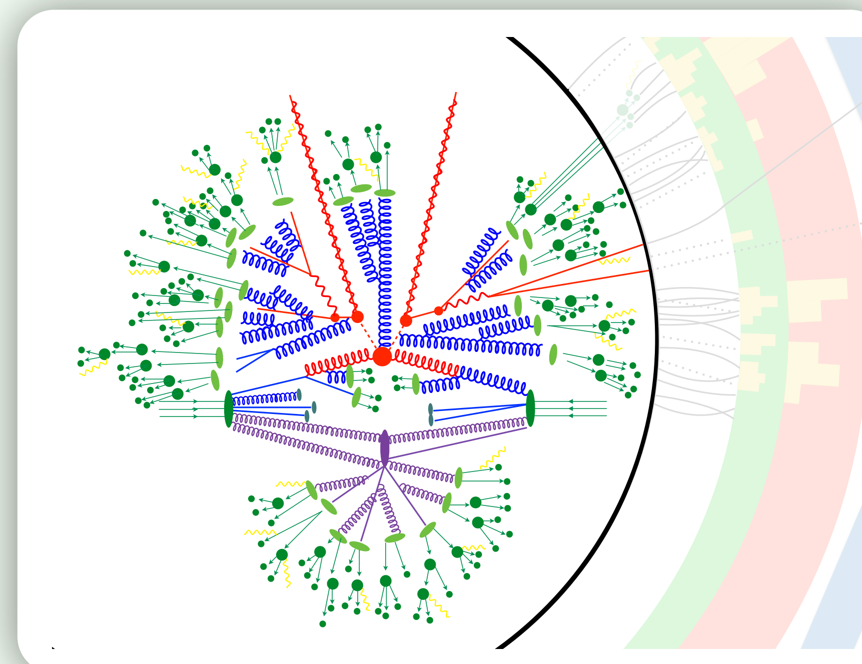
The Unfolding Challenge

Particle Level

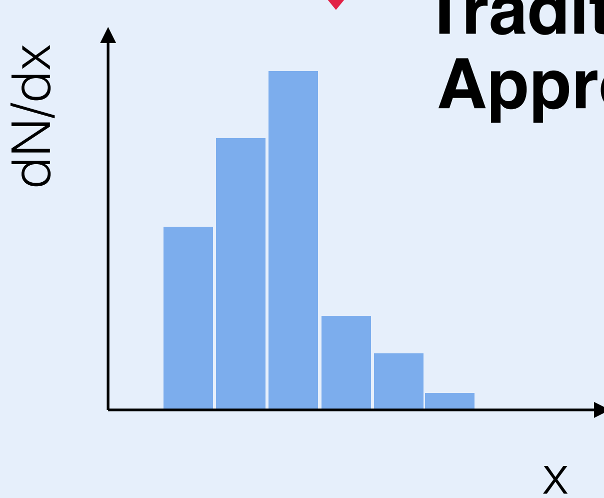
Measure this

Want this



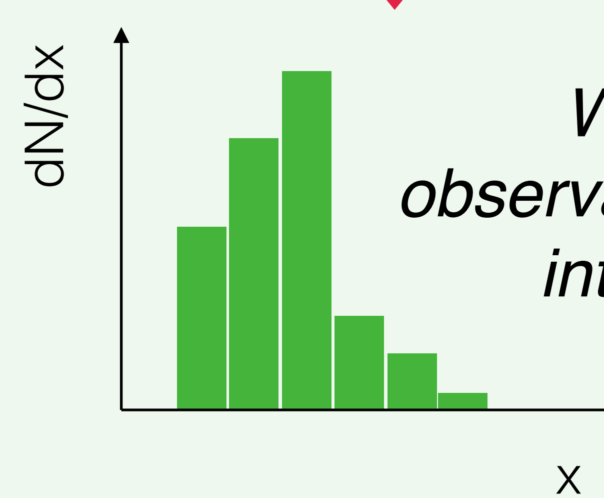


Target observable



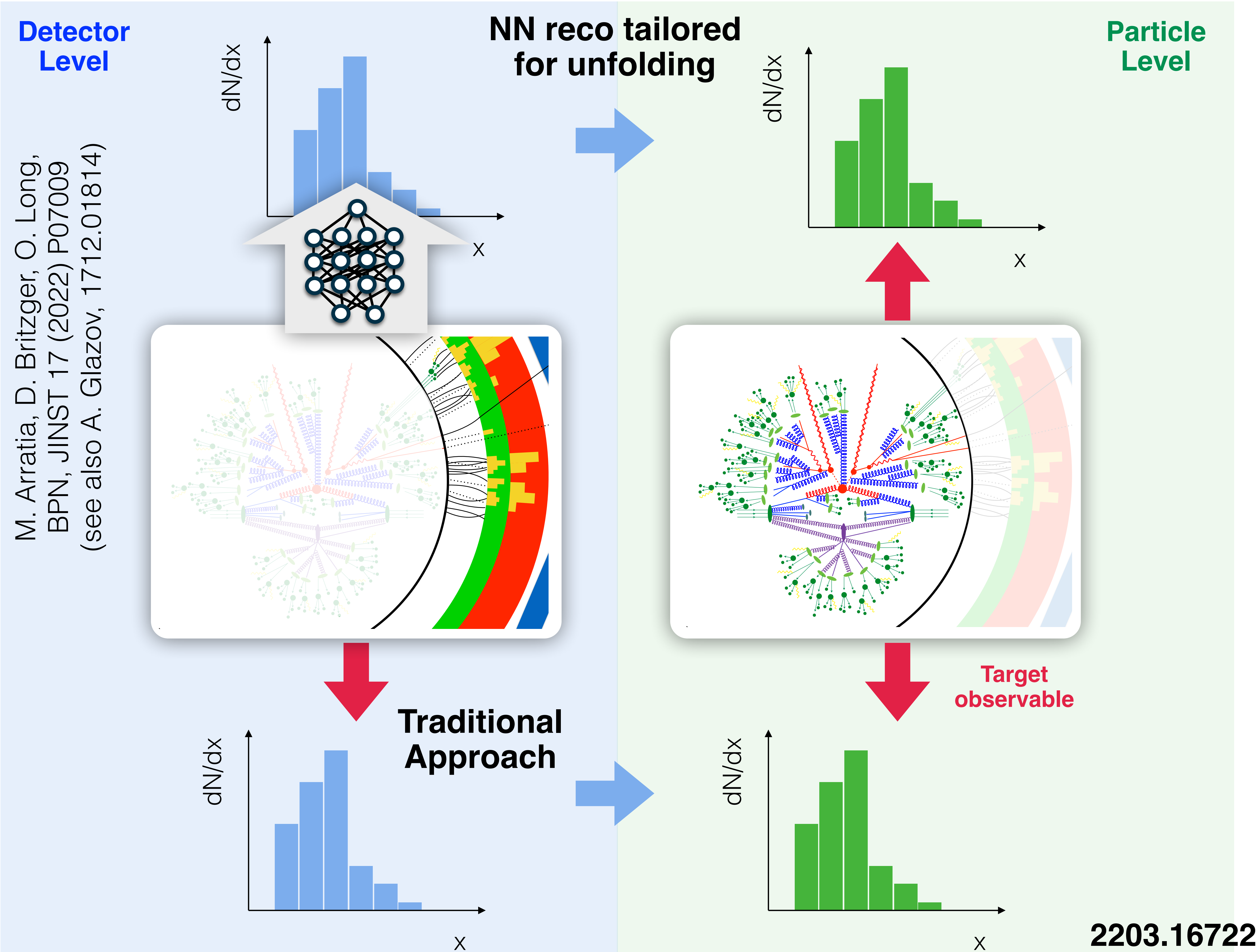
Traditional Approach

$$\hat{T} \approx R^{-1}M$$

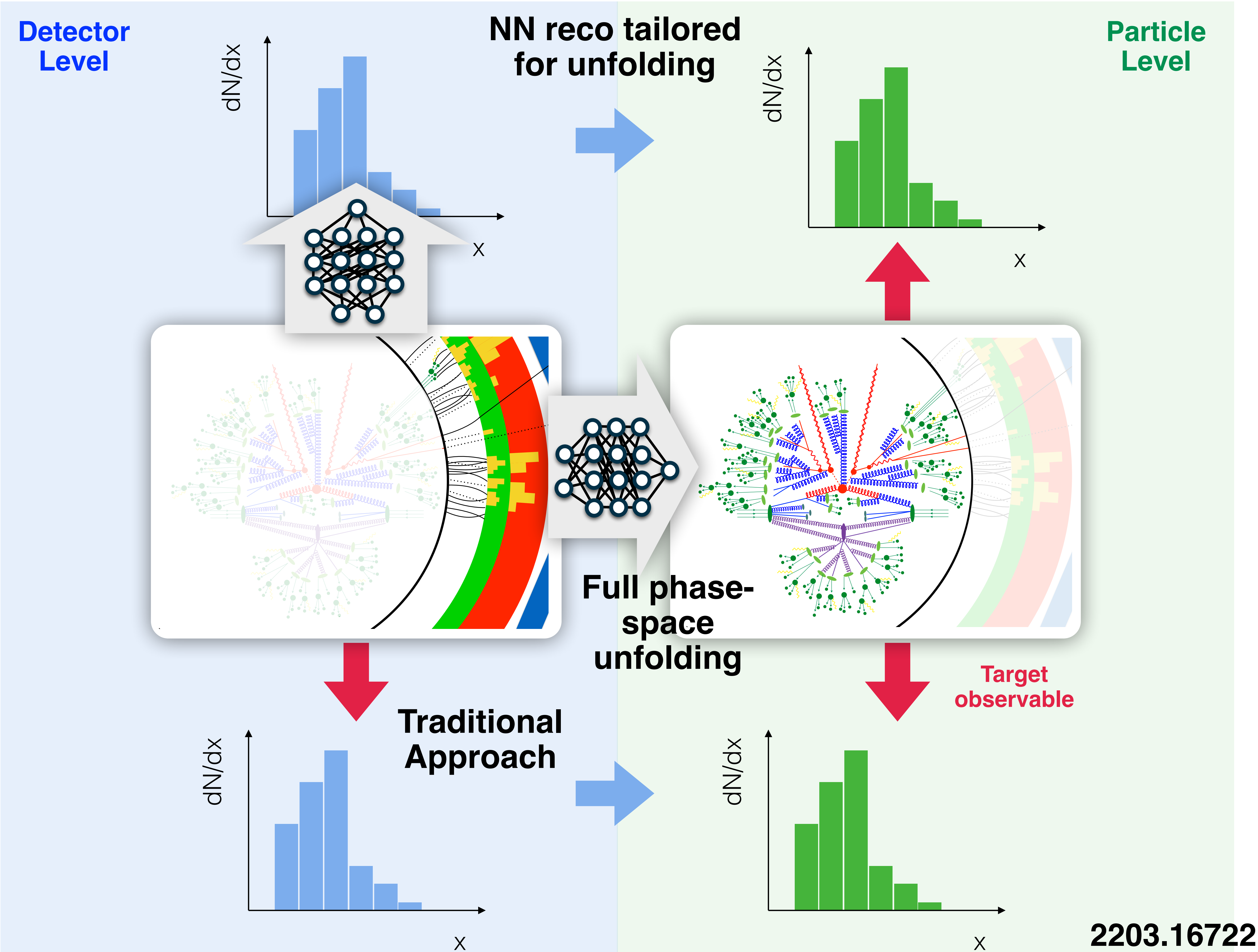


We pick O(1) observables & discretize into O(10) bins

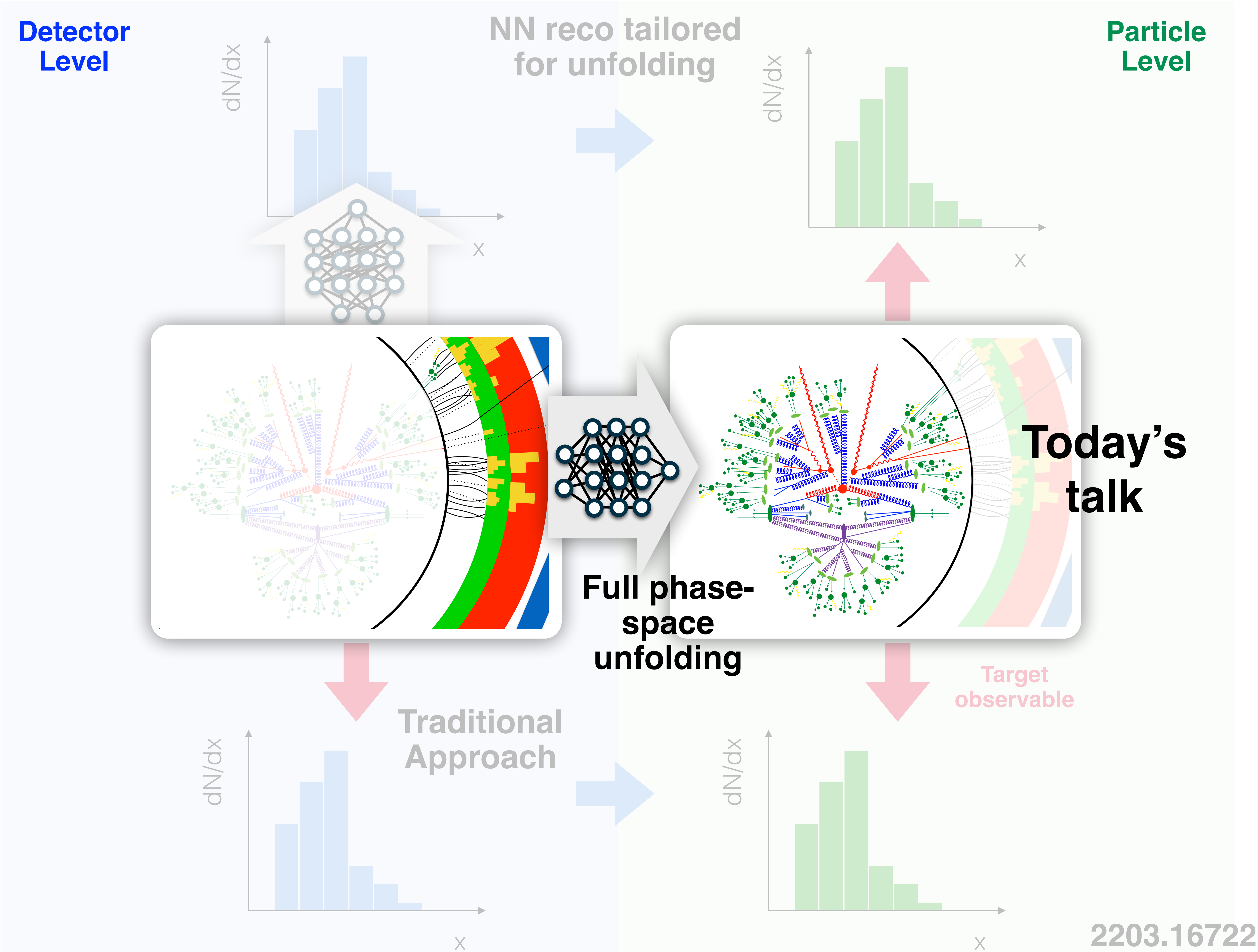
2203.16722



Detector Level
Particle Level
NN reco tailored for unfolding
Traditional Approach
Target observable
dN/dx
x
M. Arratia, D. Britzger, O. Long, BPN, JINST 17 (2022) P07009 (see also A. Glazov, 1712.01814)
2203.16722



Detector Level
NN reco tailored for unfolding
Particle Level
dN/dx
x
dN/dx
x
Full phase-space unfolding
Target observable
Traditional Approach
dN/dx
x
dN/dx
x
2203.16722


Detector Level
NN reco tailored for unfolding
Particle Level
dN/dx
x
dN/dx
x
Today's talk
Full phase-space unfolding
Target observable
Traditional Approach
dN/dx
x
dN/dx
x
2203.16722

Why unbinned (+high-dimensional)?

For a community white paper, see JINST 17 (2022) P01024, 2109.13243

Why unbinned (+high-dimensional)?

Inference-Aware Binning

Optimal binning depends on downstream task. Not possible with current setup.

What about moments?
(see also K. Desai, BPN, J. Thaler, [paper])

For a community white paper, see JINST 17 (2022) P01024, 2109.13243

Why unbinned (+high-dimensional)?

Inference-Aware Binning

Optimal binning depends on downstream task. Not possible with current setup.

What about moments?
(see also K. Desai, BPN, J. Thaler, [paper])

Derivative Measurements

With binned measurements, essentially impossible to re-use results for a function of the phase space.

For a community white paper, see JINST 17 (2022) P01024, 2109.13243

Why unbinned (+high-dimensional)?

Inference-Aware Binning

Optimal binning depends on downstream task. Not possible with current setup.

What about moments?
(see also K. Desai, BPN, J. Thaler, [paper])

Derivative Measurements

With binned measurements, essentially impossible to re-use results for a function of the phase space.

Higher Dimensions

Some phenomena can't be probed in a few dimensions.

What about observables that are not per-event?

For a community white paper, see JINST 17 (2022) P01024, 2109.13243

Landscape of Methods

Classifier-Based Methods

Learn (unfolded) data likelihood ratio w.r.t. simulation

For references, see JINST 17 (2022) P01024, 2109.13243 (and papers that cite it!)

Landscape of Methods

Classifier-Based Methods

Learn (unfolded) data likelihood ratio w.r.t. simulation

Density-Based Methods

Learn (unfolded) data probably density implicitly or explicitly.

For references, see JINST 17 (2022) P01024, 2109.13243 (and papers that cite it!)

Landscape of Methods

Classifier-Based Methods

Learn (unfolded) data likelihood ratio w.r.t. simulation

I'll focus here today because:

Learn a small correction (start close to the right answer)

&

~prior independent (if maximum likelihood)

Density-Based Methods

Learn (unfolded) data probably density implicitly or explicitly.

For references, see JINST 17 (2022) P01024, 2109.13243 (and papers that cite it!)

Landscape of Methods

Classifier-Based Methods

Learn (unfolded) data likelihood ratio w.r.t. simulation

I'll focus here today because:

Learn a small correction (start close to the right answer)

&

~prior independent (if maximum likelihood)

Density-Based Methods

Learn (unfolded) data probably density implicitly or explicitly.

I won't talk about these at all, but there has been a lot of work with GANs, VAEs, NFs, Diffusion…

GANs: K. Datta, D. Kar, D. Roy, 1806.00433; M. Bellagente, A. Butter, G. Kasieczka, T. Plehn, R. Winterhalder, SciPost Phys. 8 (2020) 070, …

VAEs: J. Howard, S. Mandt, D. Whiteson, Y. Yang, Sci. Rep. 12 (2022) 7567, …

NFs: M. Bellagente et al., SciPost Phys. 9 (2020) 074; M. Vandegar, M. Kagan, A. Wehenkel, G. Louppe, PMLR 11 (2021) 2107; M. Backes, A. Butter, M. Dunford, B. Malaescu, 2212.08674, …

Diffusion: A. Shmakov et al., 2305.10399; S. Diefenbachar, G. Liu, V. Mikuni, B. Nachman, W. Nie, 2308.12351

For references, see JINST 17 (2022) P01024, 2109.13243 (and papers that cite it!)

Landscape of Methods

Classifier-Based Methods

Learn (unfolded) data likelihood ratio w.r.t. simulation

I'll focus here today because:

*Learn a small correction**
(start close to the right answer)

&

~prior independent
(if maximum likelihood)

Density-Based Methods

Learn (unfolded) data probably density implicitly or explicitly.

I won't talk about these at all, but there has been a lot of work with GANs, VAEs, NFs, Diffusion…

GANs: K. Datta, D. Kar, D. Roy, 1806.00433; M. Bellagente, A. Butter, G. Kasieczka, T. Plehn, R. Winterhalder, SciPost Phys. 8 (2020) 070, …

VAEs: J. Howard, S. Mandt, D. Whiteson, Y. Yang, Sci. Rep. 12 (2022) 7567, …

NFs: M. Bellagente et al., SciPost Phys. 9 (2020) 074; M. Vandegar, M. Kagan, A. Wehenkel, G. Louppe, PMLR 11 (2021) 2107; M. Backes, A. Butter, M. Dunford, B. Malaescu, 2212.08674, …

Diffusion: A. Shmakov et al., 2305.10399; S. Diefenbachar, G. Liu, V. Mikuni, B. Nachman, W. Nie, 2308.12351*

For references, see JINST 17 (2022) P01024, 2109.13243 (and papers that cite it!)

Landscape of Methods

Classifier-Based Methods

Learn (unfolded) data likelihood ratio w.r.t. simulation

I'll focus here today because:

Learn a small correction (start close to the right answer)

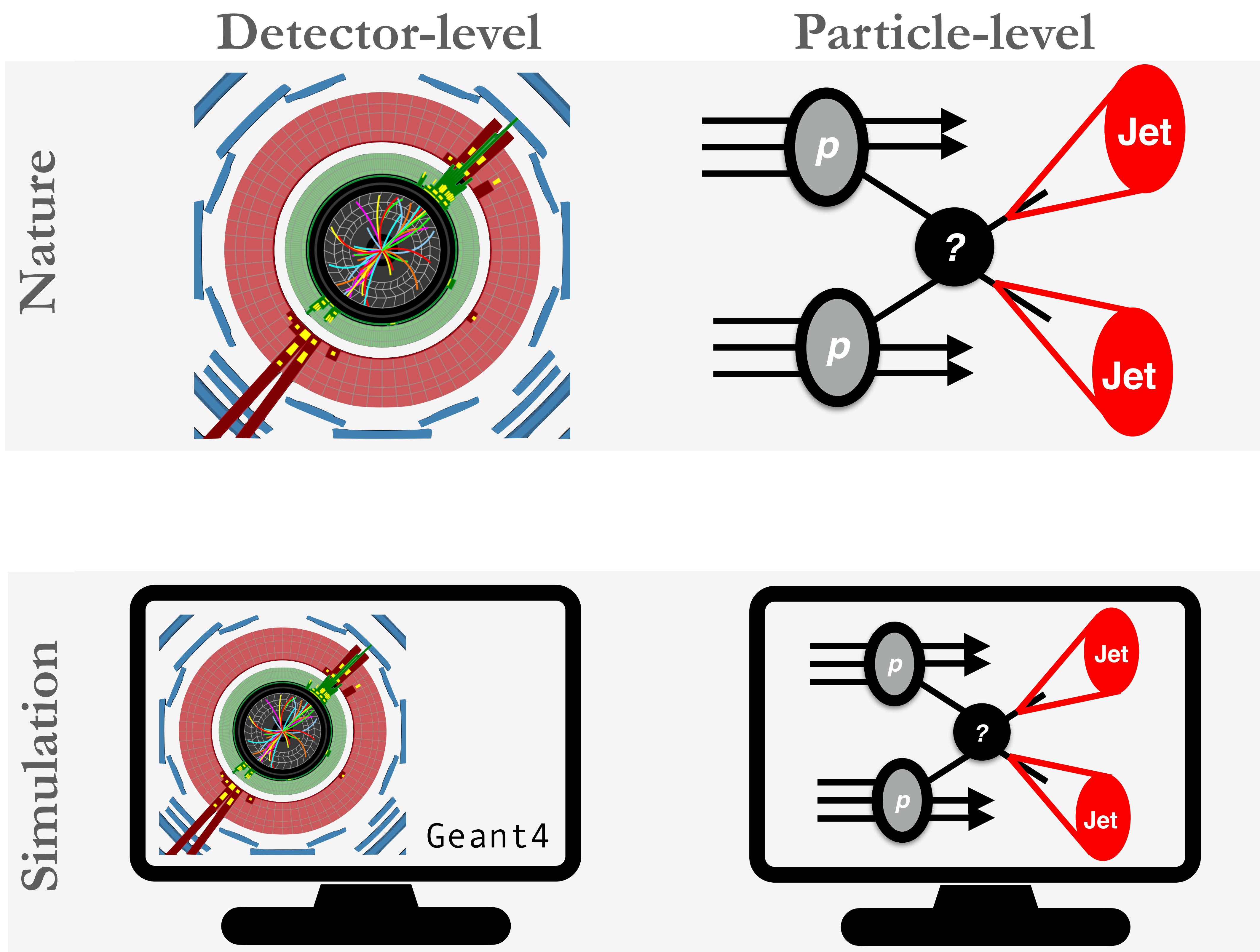
&

~prior independent (if maximum likelihood)

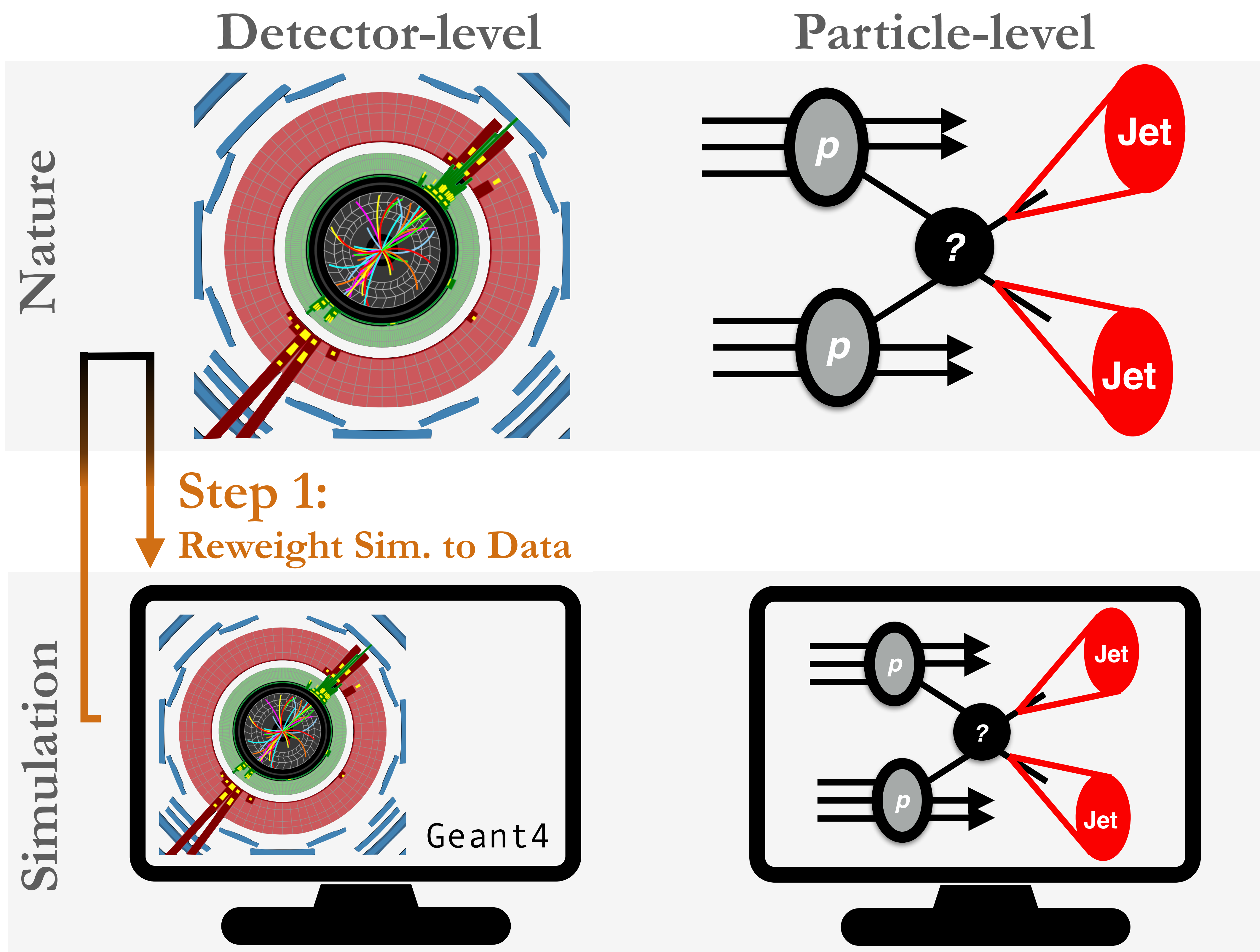
My focus will be on a method called **OmniFold**.

For references, see JINST 17 (2022) P01024, 2109.13243 (and papers that cite it!)

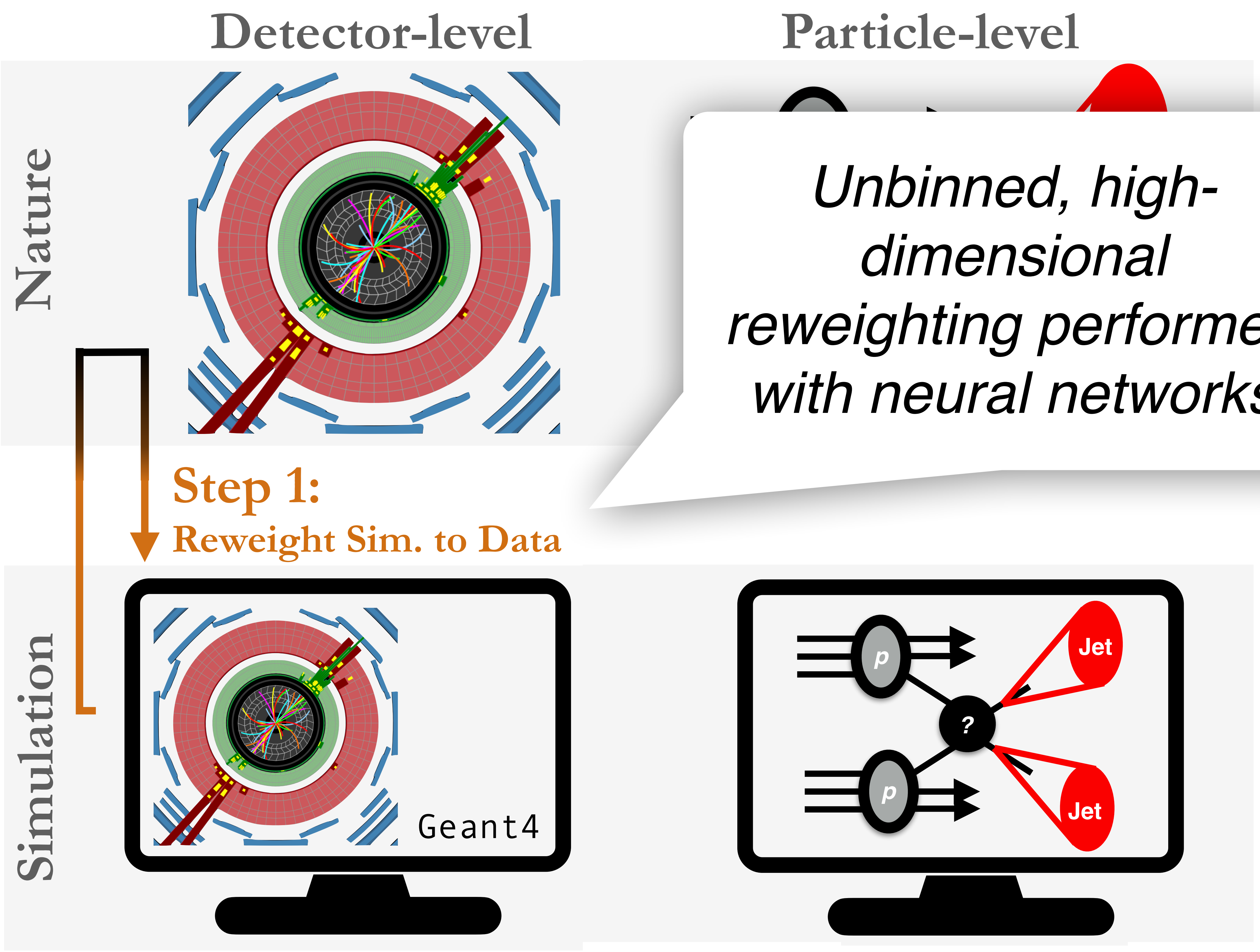
A brief introduction to OmniFold



A brief introduction to OmniFold

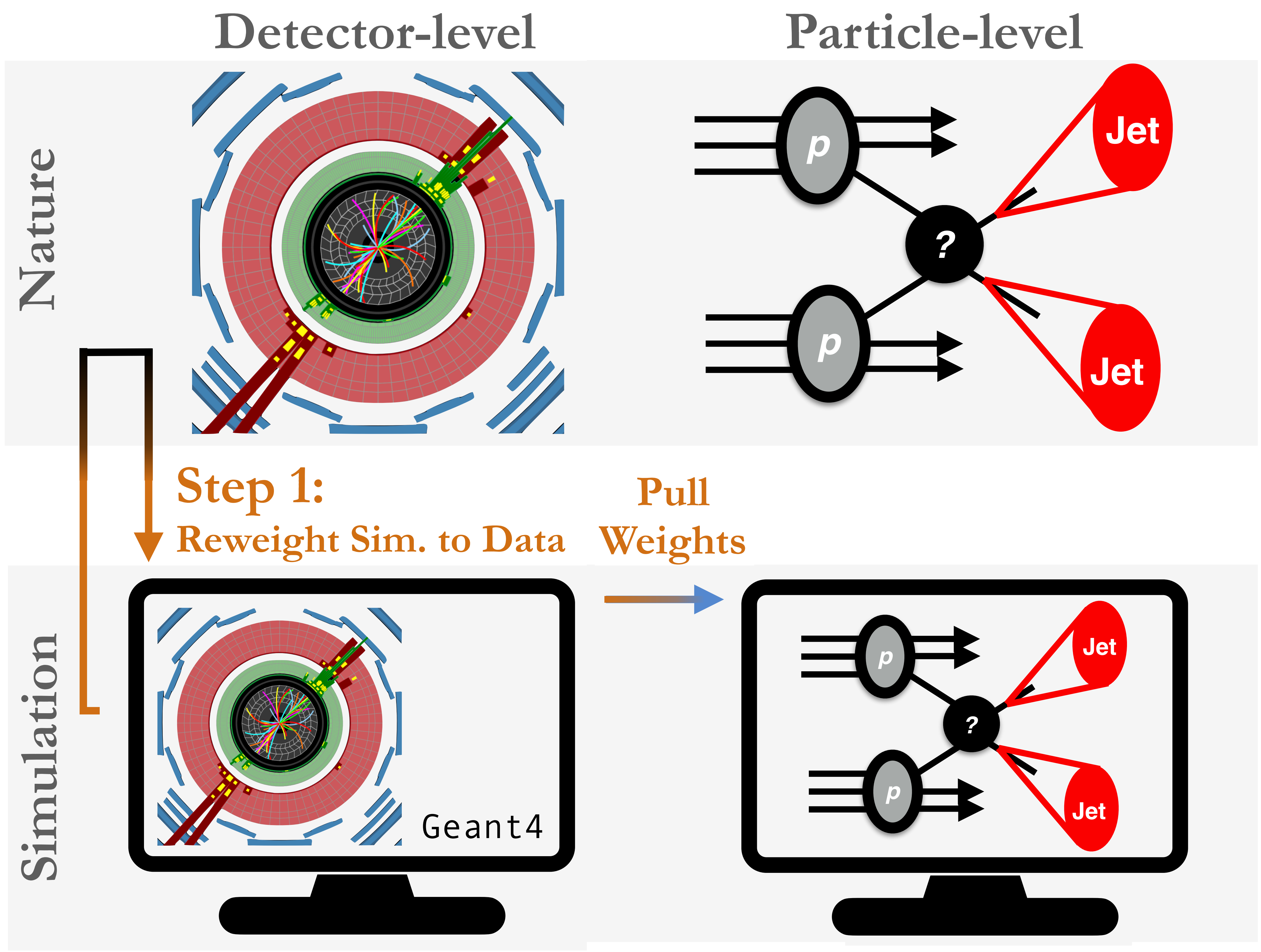


A brief introduction to OmniFold

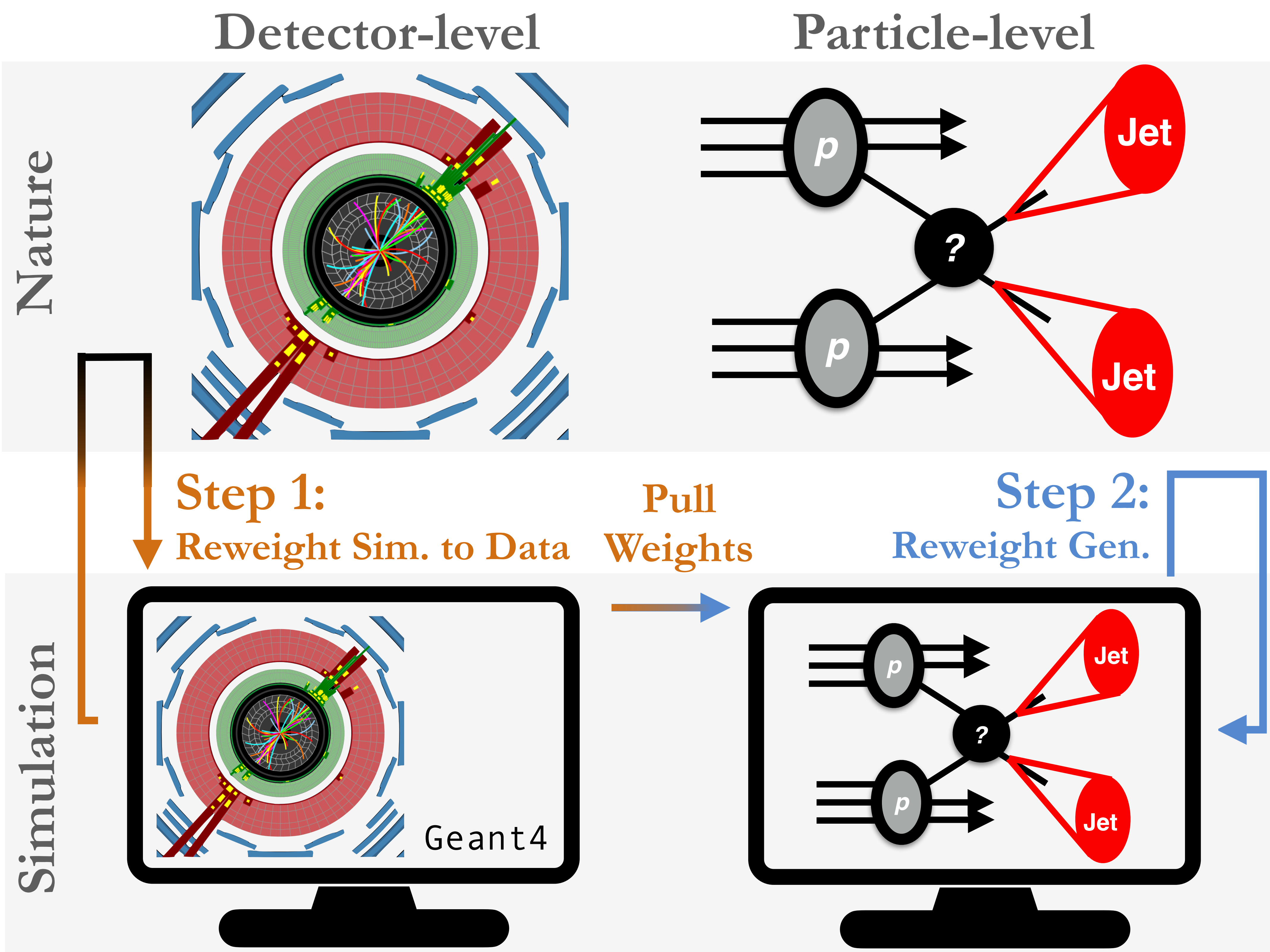


Unbinned, high-dimensional reweighting performed with neural networks

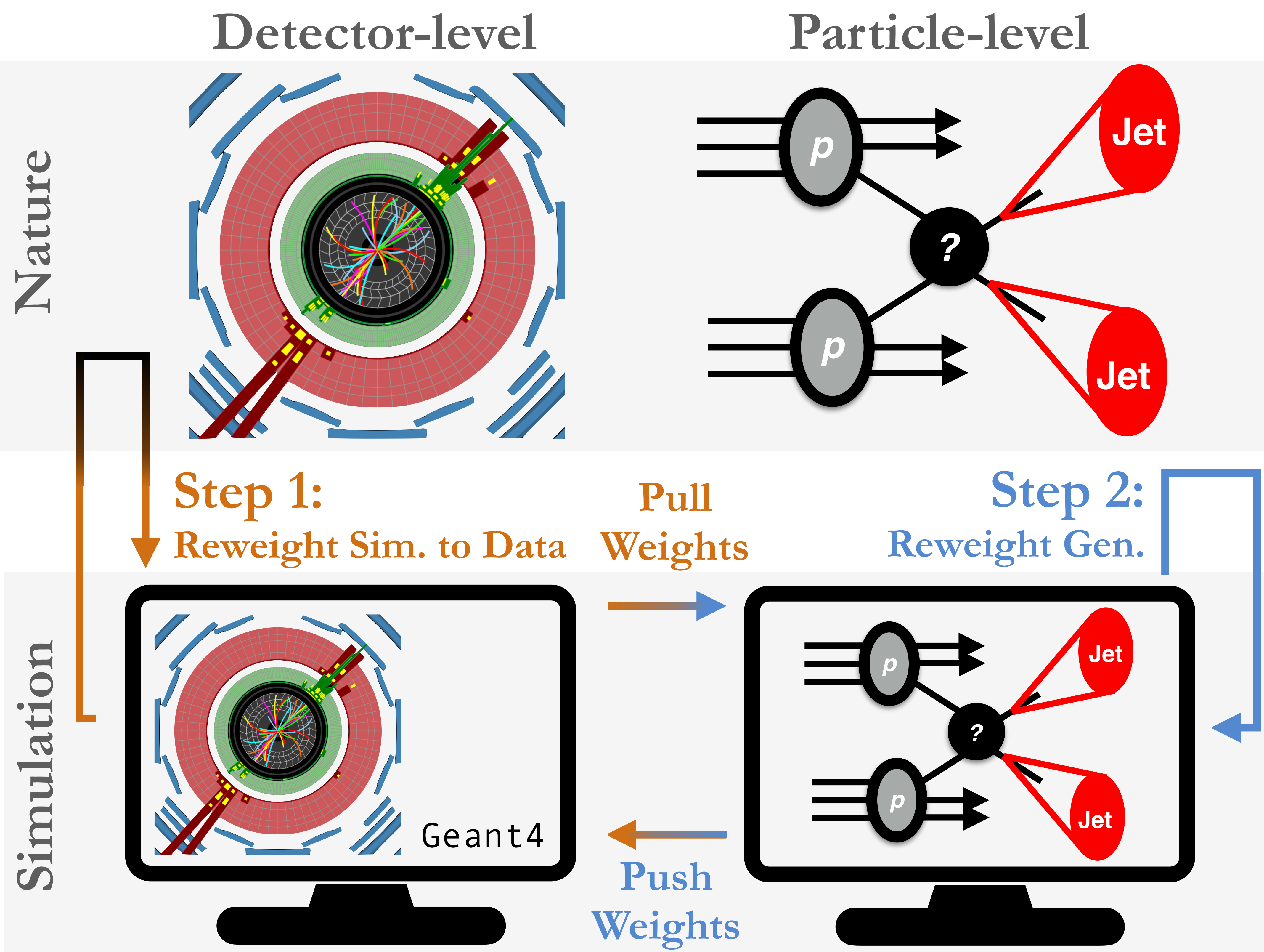
A brief introduction to OmniFold



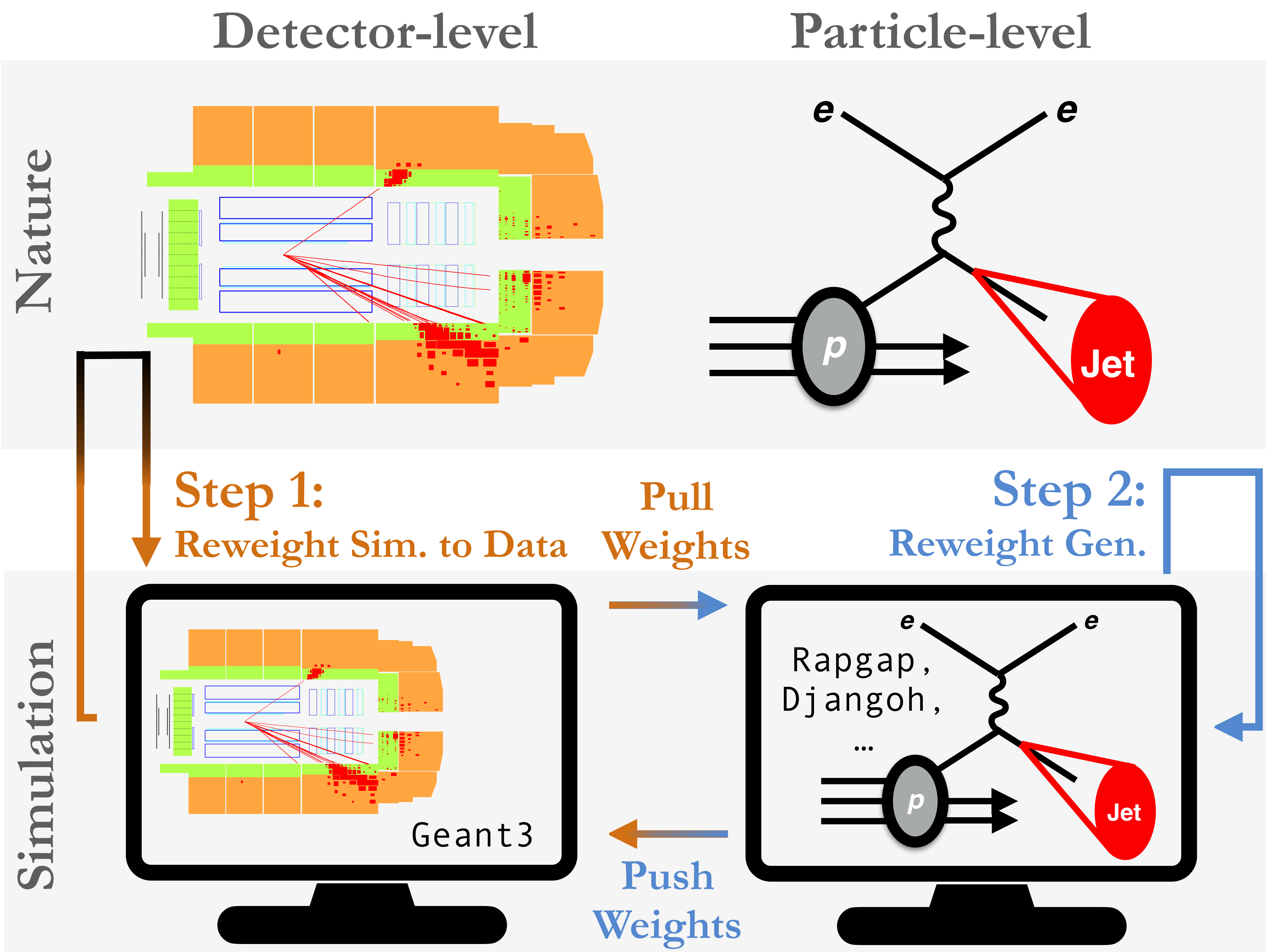
A brief introduction to OmniFold



A brief introduction to OmniFold

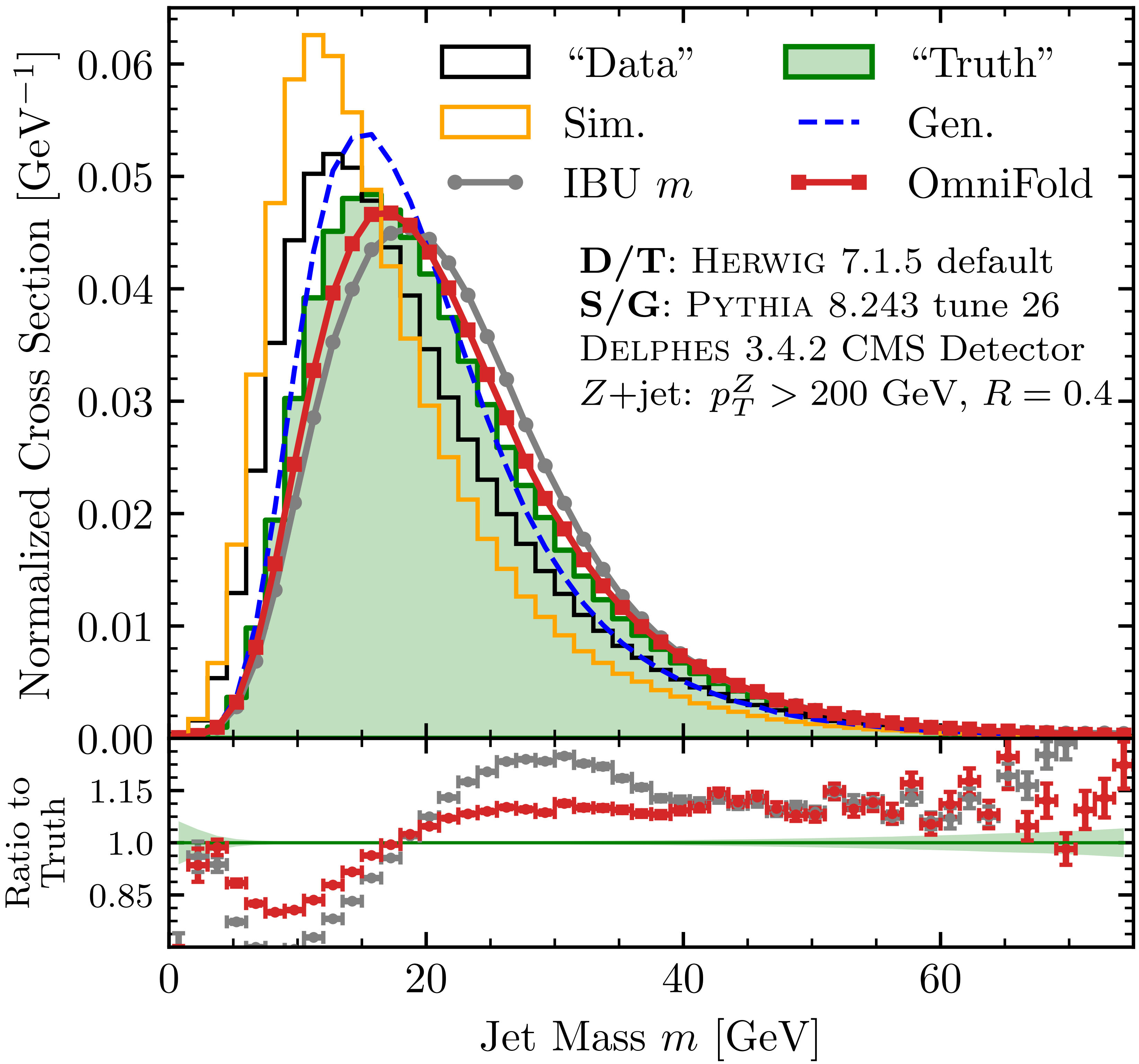


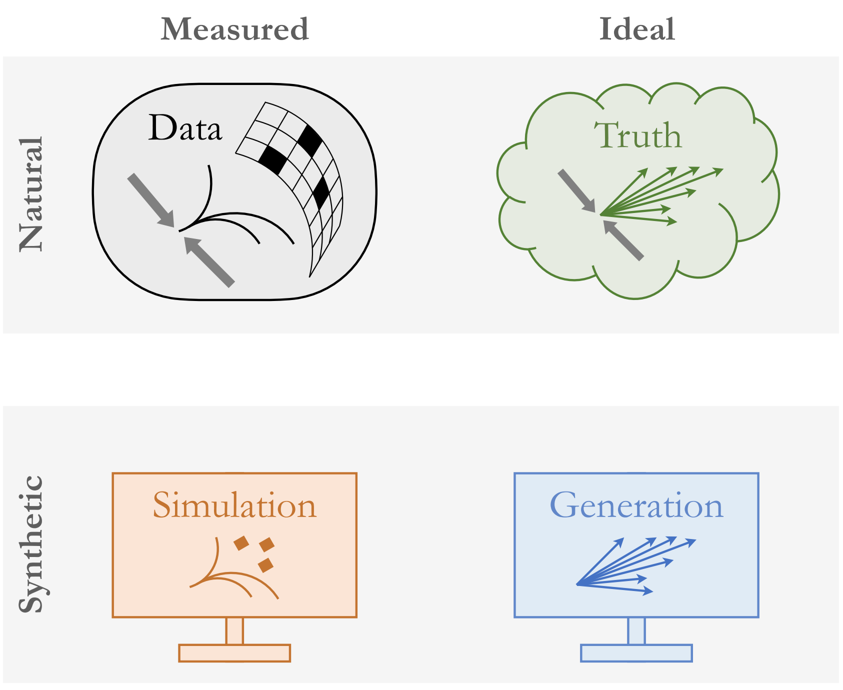
A brief introduction to OmniFold



Full phase-space unfolding

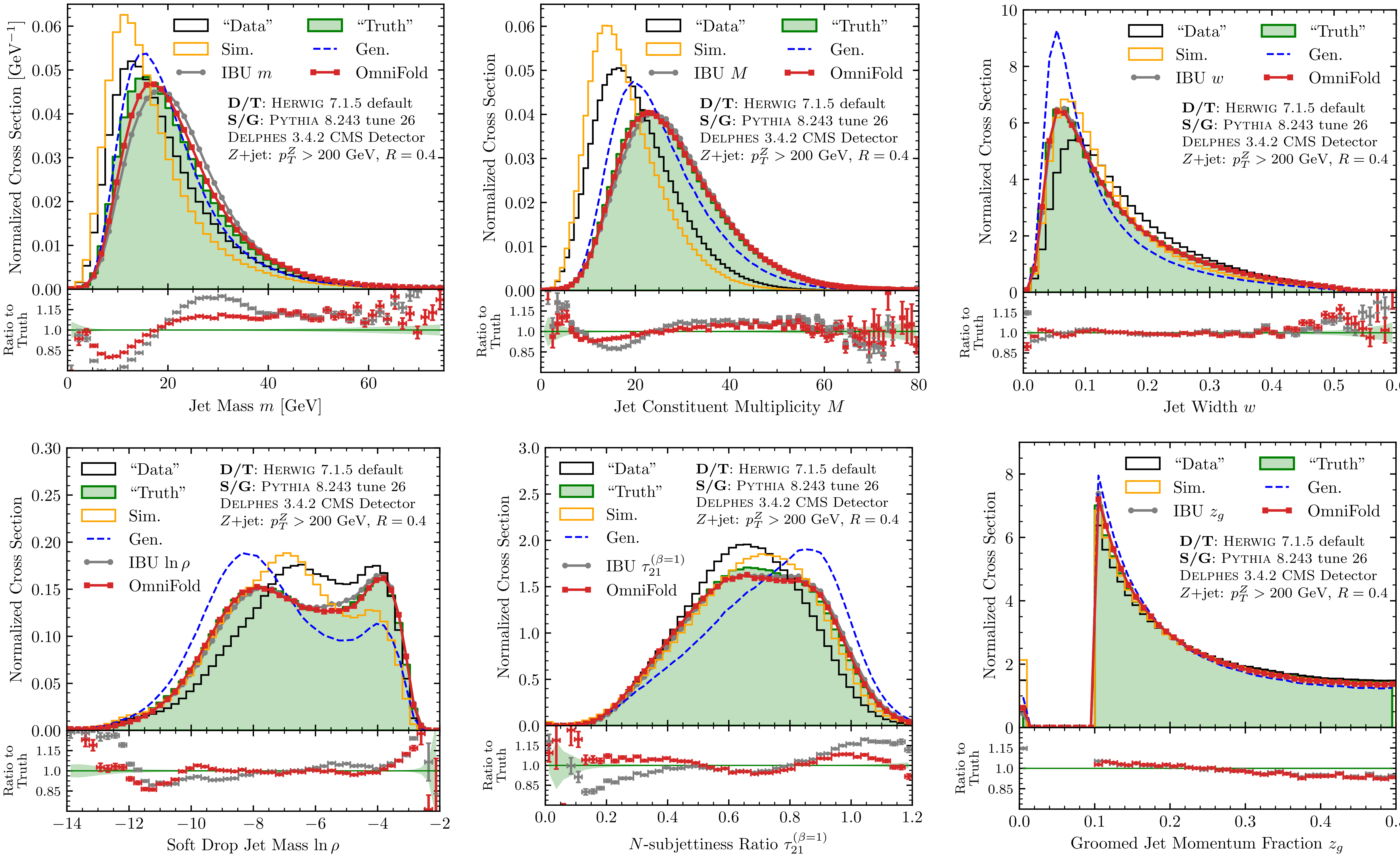
A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001





Full phase-space unfolding

A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001



Full phase-space unfolding

OmniFold is:

- *Unbinned*
- *Maximum likelihood**
- *Improves the resolution from correlations with detector response*

**when binned, OmniFold converges to Lucy-Richardson (aka Iterative Bayesian Unfolding)*

*In fact, OmniFold can also work on low-level inputs (e.g. energy flow particles). In that case, you can construct observables **after** the measurement.*

Some technical details

Please ask if you are interested, but briefly, OmniFold…

- Can accommodate backgrounds (unbinned) via neural positive reweighing
- Can accommodate acceptance effects
- Has a number of choices for how to update weights and/or keep track of acceptance effects

https://github.com/hep-lbdl/OmniFold

See A. Andreassen et al., ICLR SimDL for details [https://simdl.github.io/files/12.pdf]

First Results

I’ll now spend a ~1 minute flashing the first unbinned measurement results

There is no time to give the physics content justice, so I’ll be brief, but please let me know if you have any questions!

Results from H1, LHCb, STAR, …

New methods for unbinned unfolding are here! We should be ready to use them also for BSM!

DESY 21-130, ISSN 0418-9833

Measurement of lepton-jet correlation in deep-inelastic scatterin[g] with the H1 detector using machine learning for unfolding

(The H1 Collaboration)

arXiv:2108.12376v2 [hep-ex] 1 Apr 2022

EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH (CERN)

CERN-EP-2022-161
LHCb-PAPER-2022-013
August 25, 2022

Multidifferential study of identified charged hadron distributions in Z-tagged jets in proton-proton collisions at $\sqrt{s}$ =13 TeV

Abstract

Jet fragmentation functions are measured for the first time in proton-proton collisions for charged pions, kaons, and protons within jets recoiling against a Z boson. The charged-hadron distributions are studied longitudinally and transversely to the jet direction for jets with transverse momentum $20 < p_T < 100$ GeV and in the pseudorapidity range $2.5 < \eta < 4$. The data sample was collected with the LHCb experiment at a center-of-mass energy of 13 TeV, corresponding to an integrated luminosity of 1.64 fb^{-1}. Triple differential distributions as a function of the hadron longitudinal momentum fraction, hadron transverse momentum, and jet transverse momentum are also measured for the first time. This helps constrain transverse-momentum-dependent fragmentation functions. Differences in the shapes and magnitudes of the measured distributions for the different hadron species provide insights into the hadronization process for jets predominantly initiated by light quarks.

Submitted to Phys. Rev. D Letter

arXiv:2208.11691v1 [hep-ex] 24 Aug 2022

[U]nbinned Deep Learning Jet Substructure Measurement in High Q^2 ep collisions at HERA

July 19, 2023

Measurement of CollinearDrop jet mass and its correlation with SoftDrop groomed jet substructure observables in $\sqrt{s} = 200$ GeV pp collisions by STAR

Youqi Song (Wright Laboratory, Yale University)
on behalf of the STAR Collaboration

Jet substructure variables aim to reveal details of the parton fragmentation and hadronization processes that create a jet. By removing collinear radiation while maintaining the soft radiation components, one can construct CollinearDrop jet observables, which have enhanced sensitivity to the soft phase space within jets. We present a CollinearDrop jet measurement, corrected for detector effects with a machine learning method, MultiFold, and its correlation with groomed jet observables, in pp collisions at $\sqrt{s} = 200$ GeV at STAR. We demonstrate that the population of jets with a large non-perturbative contribution can be significantly enhanced by selecting on higher CollinearDrop jet mass fractions. In addition, we observe an anti-correlation between the amount of grooming and the angular scale of the first hard splitting of the jet.

PRESENTED AT

DIS2023: XXX International Workshop on Deep-Inelastic Scattering and Related Subjects,
Michigan State University, USA, 27-31 March 2023

arXiv:2307.07718v2 [nucl-ex] 18 Jul 2023

+CMS open data study

Future + challenges

So far, OmniFold seems to work as designed!
Exciting to see where this will take us.

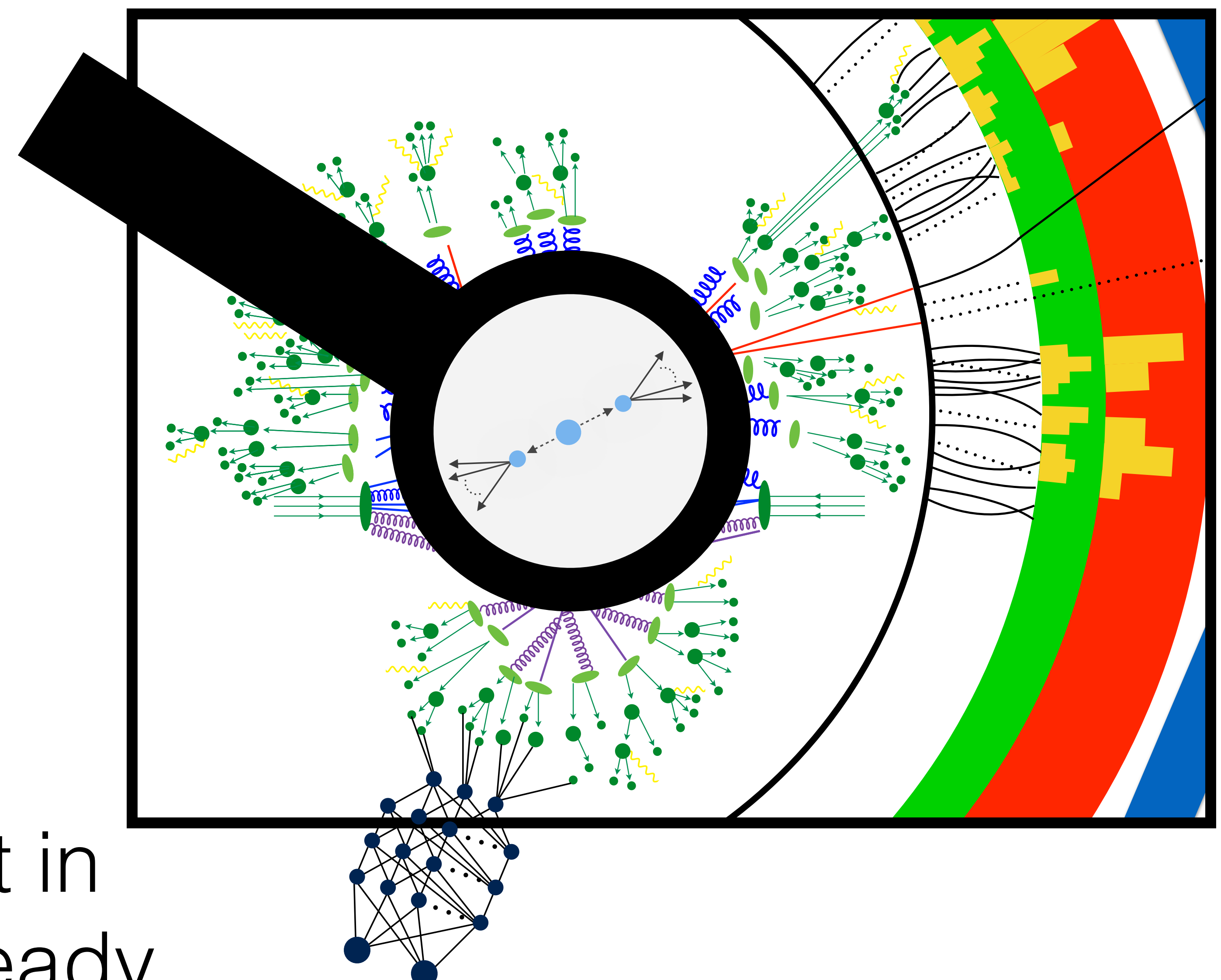
There are still some challenges we need to overcome:

- OmniFold is computationally expensive (need to train many networks, especially with ensembling to reach precision)
- How to publish an unbinned result? (all results so far are presented as binned) - see 2109.13243. Breaks HEPData!
- Modeling/closure uncertainties in high dimensions (not a new problem, but perhaps more acute)
- What about profiling? See 2302.05390 for a partial solution.

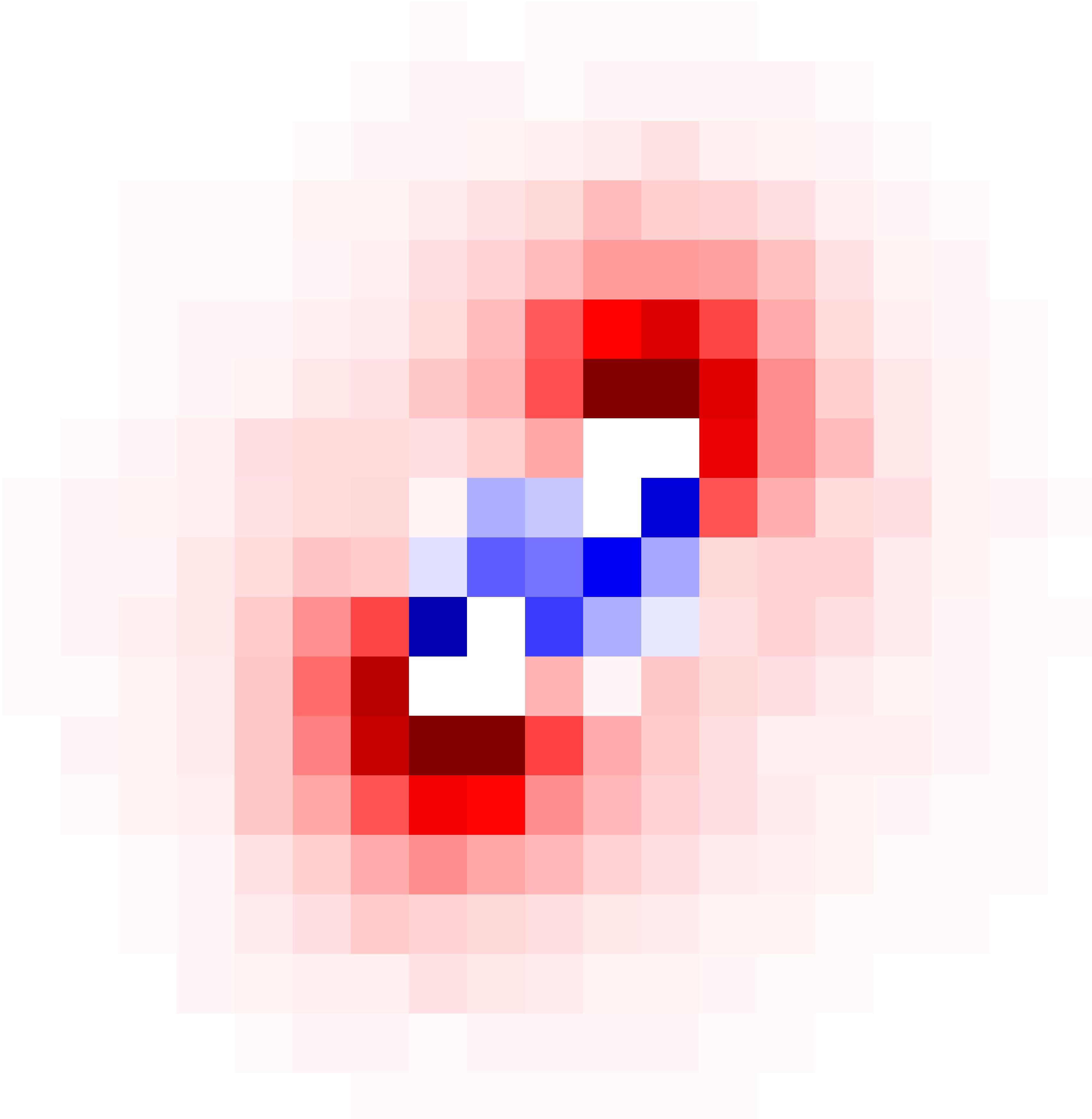
Conclusions and Outlook

A **new measurement paradigm** is possible, enabled by ML-based unfolding methods

We can analyze our data **holistically** and **future-proof** it using unbinned techniques



More R&D is required, but in parallel, these tools are already starting to **deliver science results**!



Fin.

Reweighting

How do to the reweighting without binning?

Reweighting

How do to the reweighting without binning?

dataset 1: sampled from ***p(x)***
dataset 2: sampled from ***q(x)***

Create weights $\boldsymbol{w(x) = q(x)/p(x)}$ so that when dataset 1 is weighted by ***w***, it is statistically identical to dataset 2.

What if we don't (and can't easily) know ***q*** and ***p***?

(and don't want to estimate them by binning)

Classification for reweighting

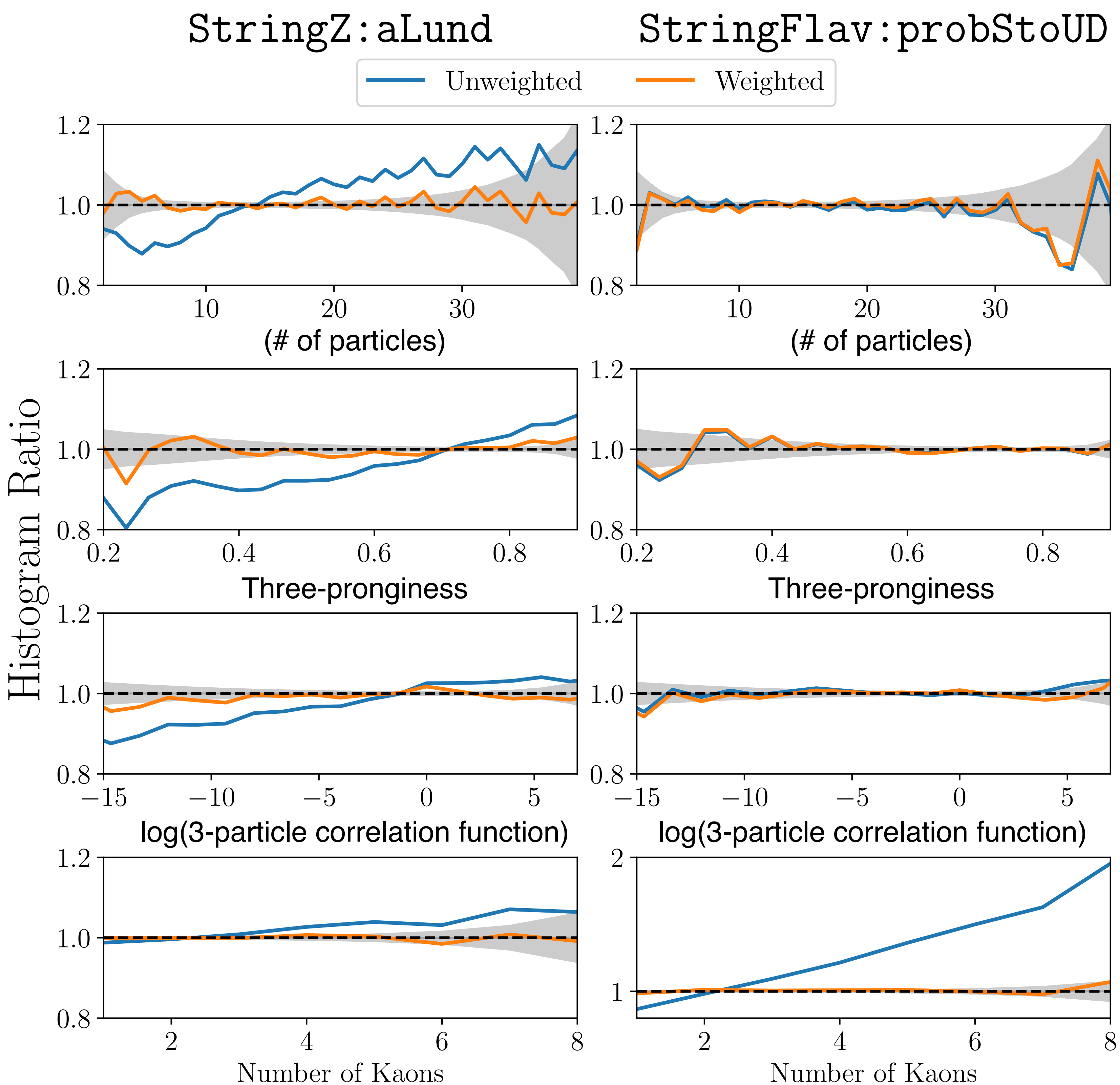
Fact: Neutral networks learn to approximate the likelihood ratio $= q(x)/p(x)$

(or something monotonically related to it in a known way)

Solution: train a neural network to distinguish the two datasets!

This turns the problem of **density estimation** (**hard**) into a problem of **classification** (**easy**)

Neural reweighing: works very well!



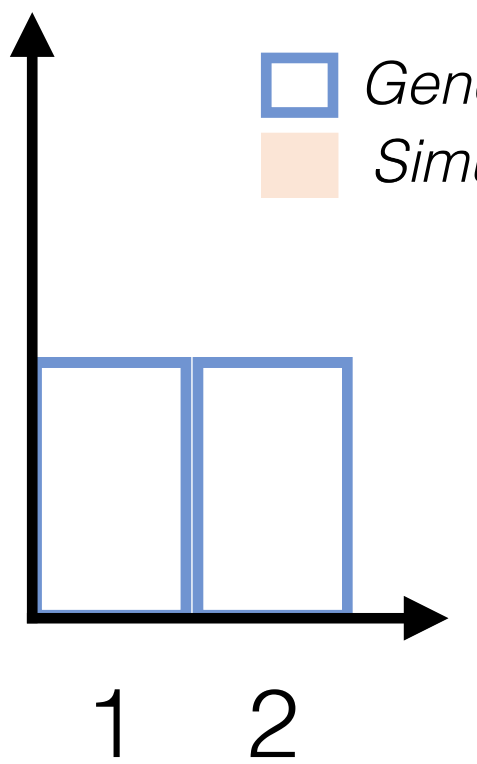
Full phase-space reweighing using simulated e^+e^-

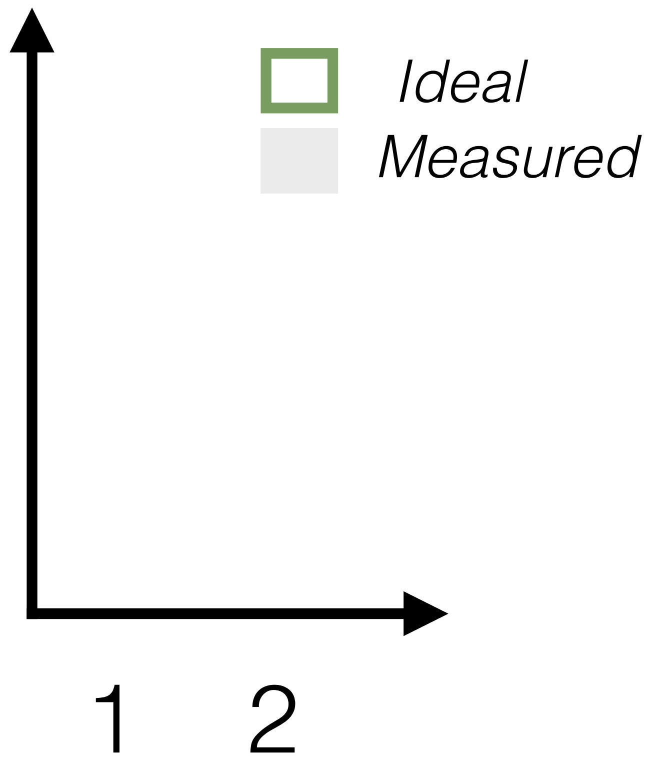
Works even when the differences are **small** (left) or **localized** (right).

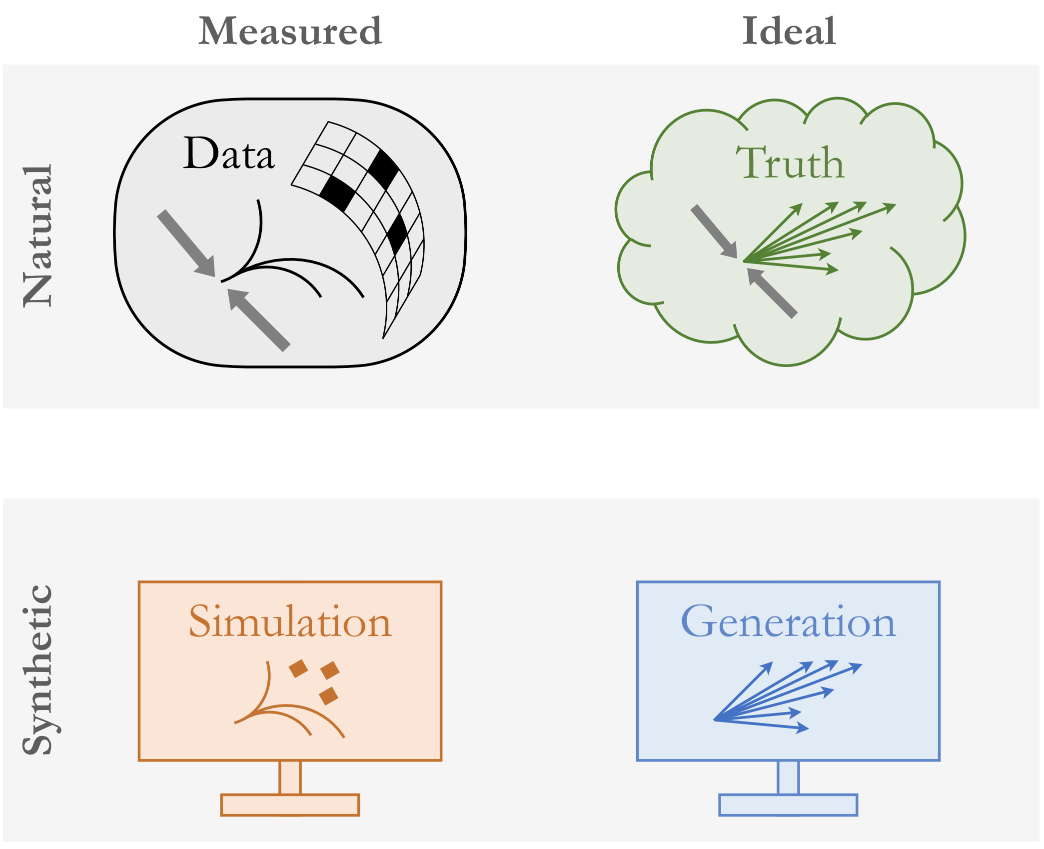
These are histogram ratios for a series of one-dimensional observables

A. Andreassen, BPN, PRD 101 (2020) 091901

Unfold by iterating: OmniFold

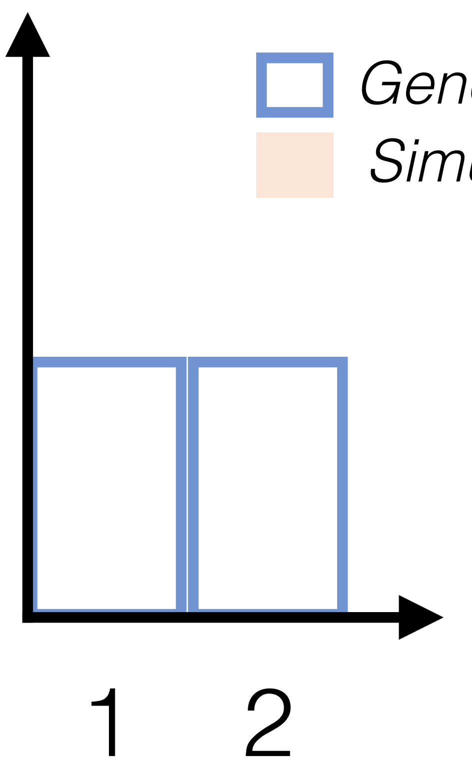


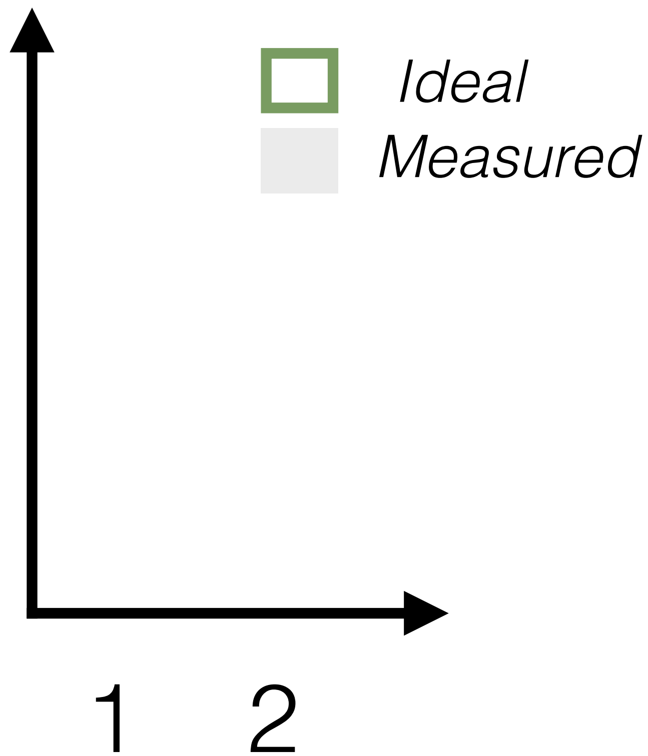




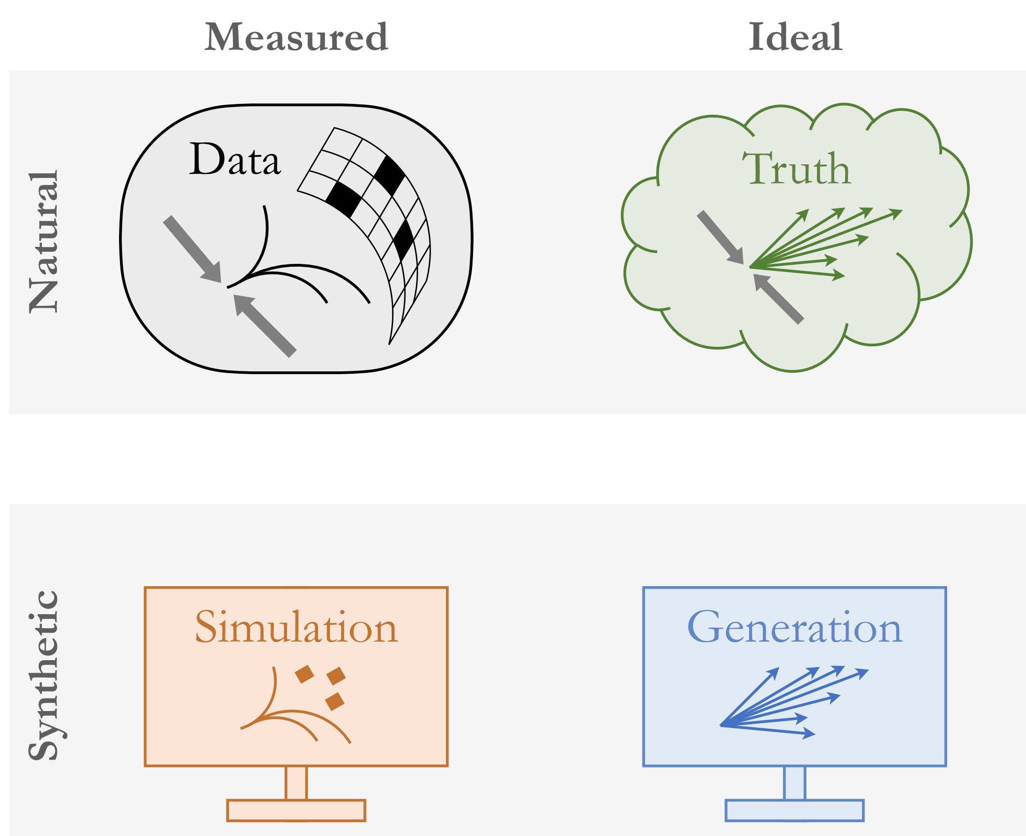
A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001

Unfold by iterating: OmniFold



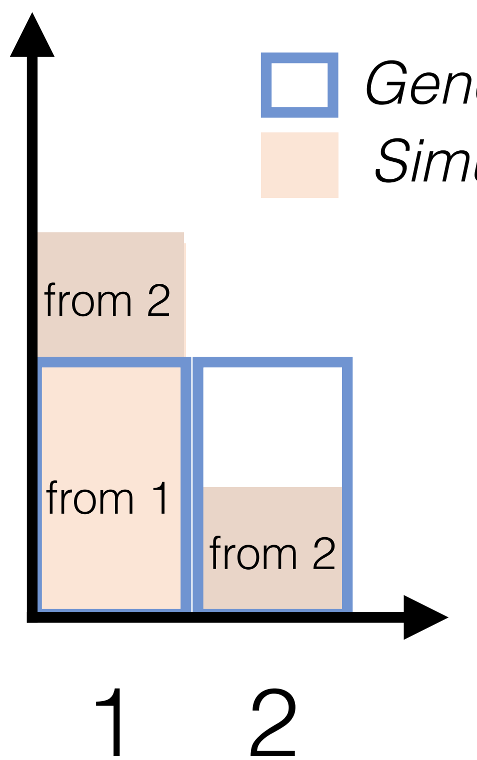


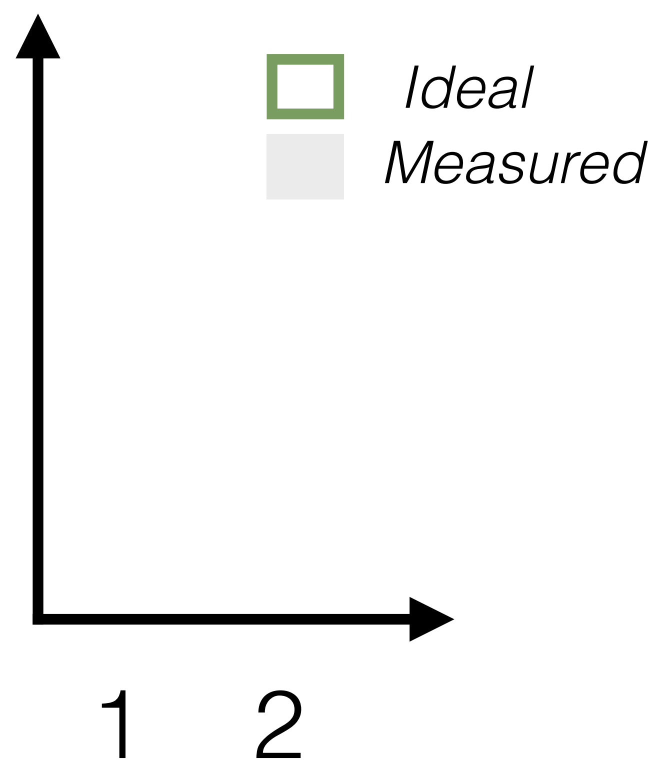
Measured \ Ideal	1	2
2	0%	50%
1	100%	50%



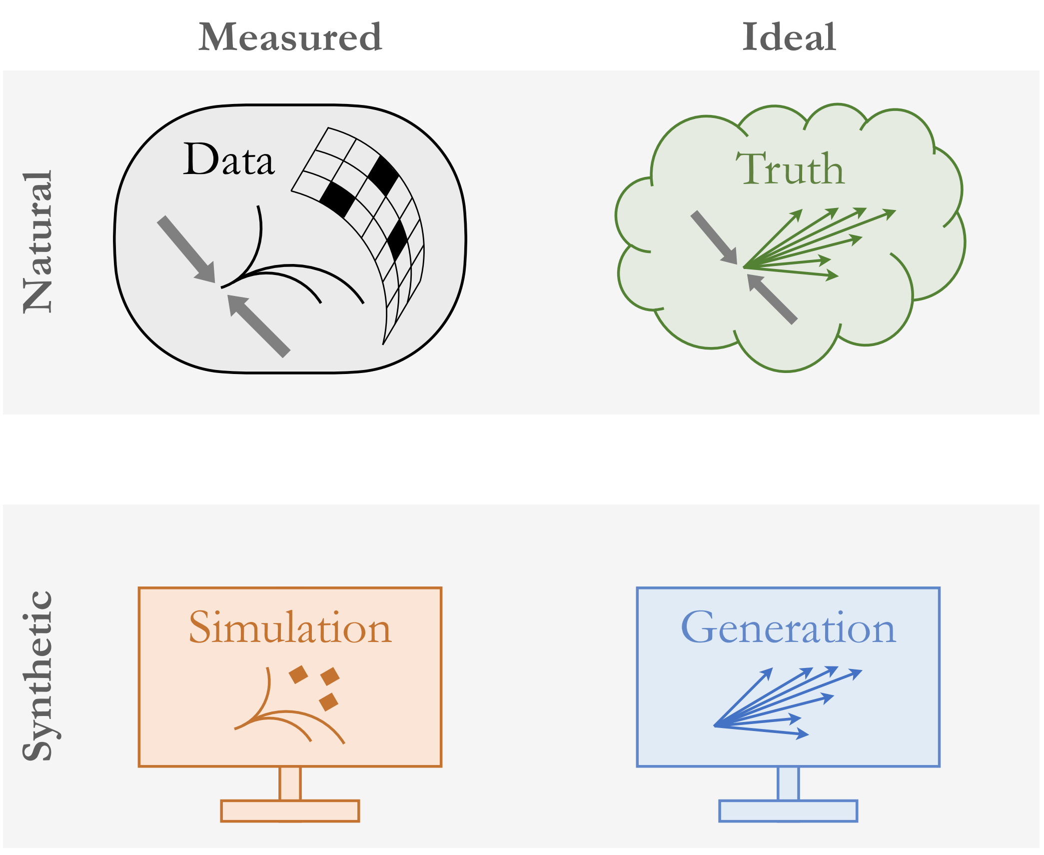
A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001

Unfold by iterating: OmniFold



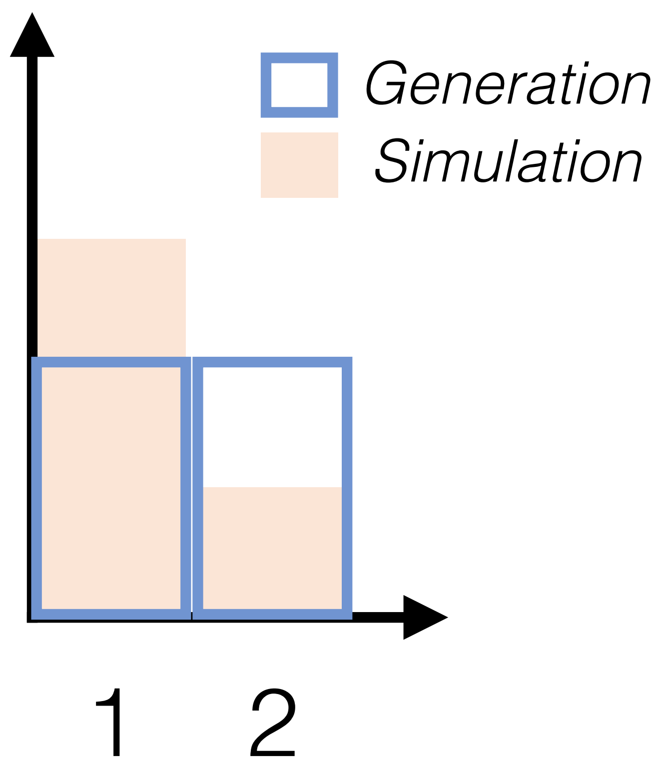


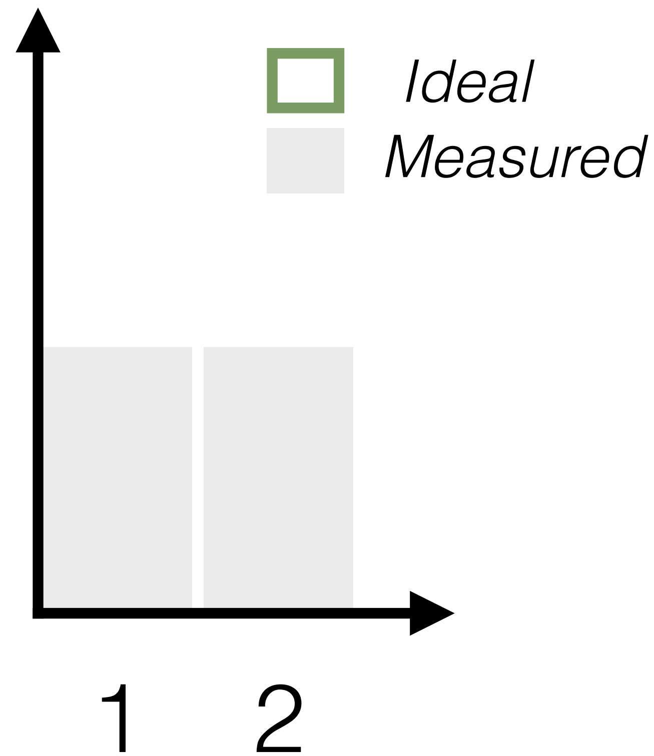
Measured \ Ideal	1	2
2	0%	50%
1	100%	50%



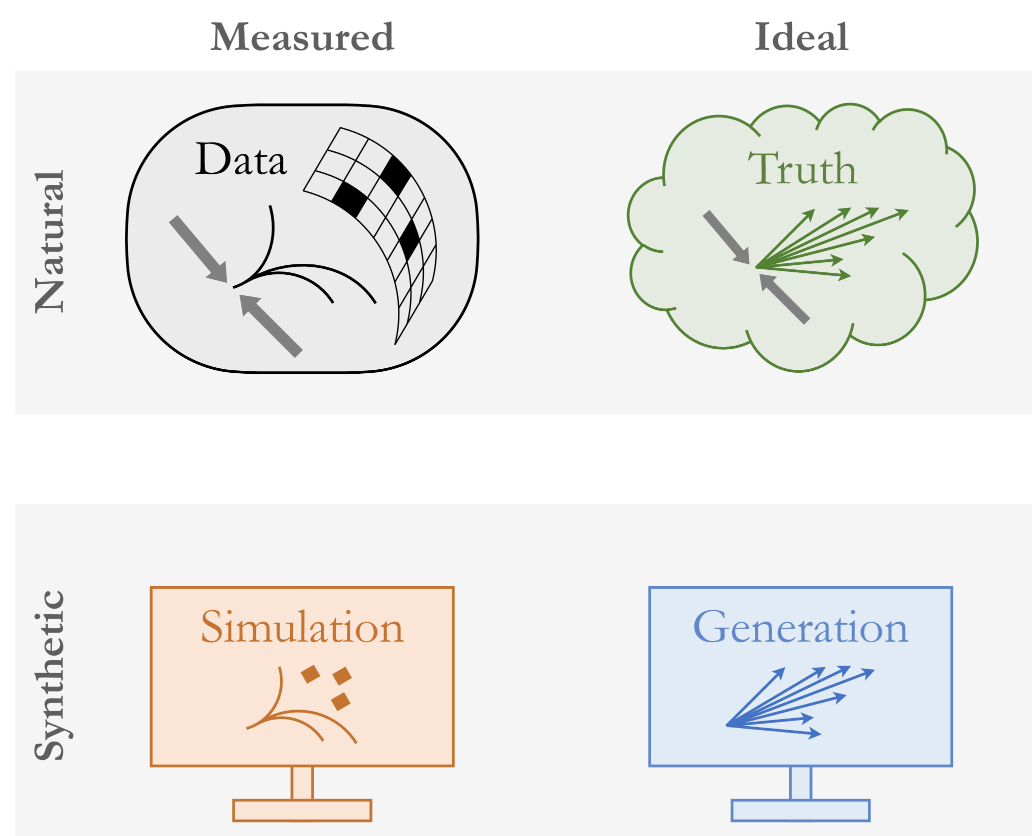
A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001

Unfold by iterating: OmniFold



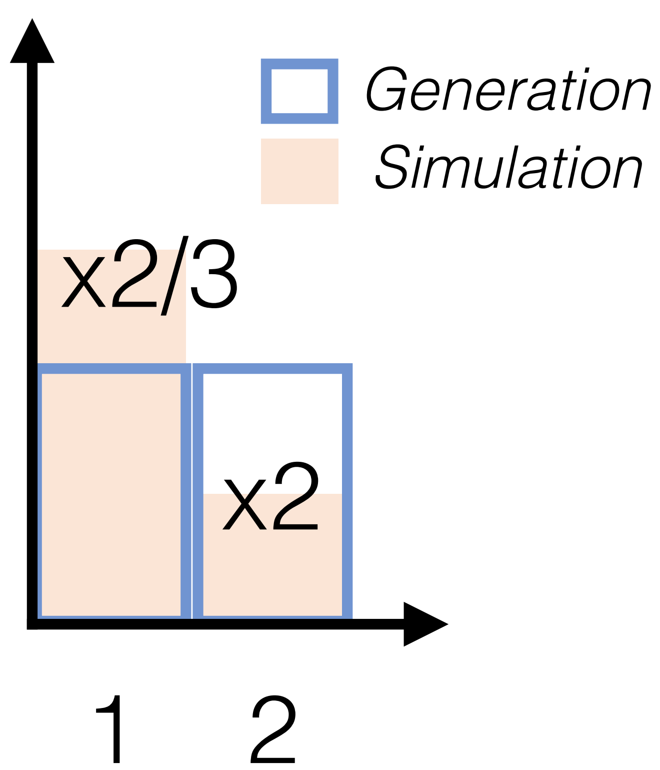


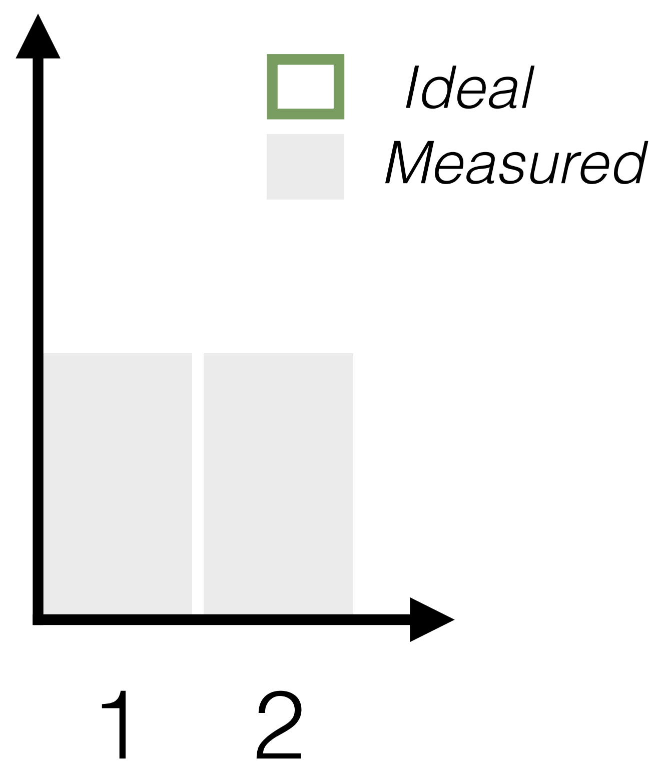
Measured \ Ideal	1	2
2	0%	50%
1	100%	50%



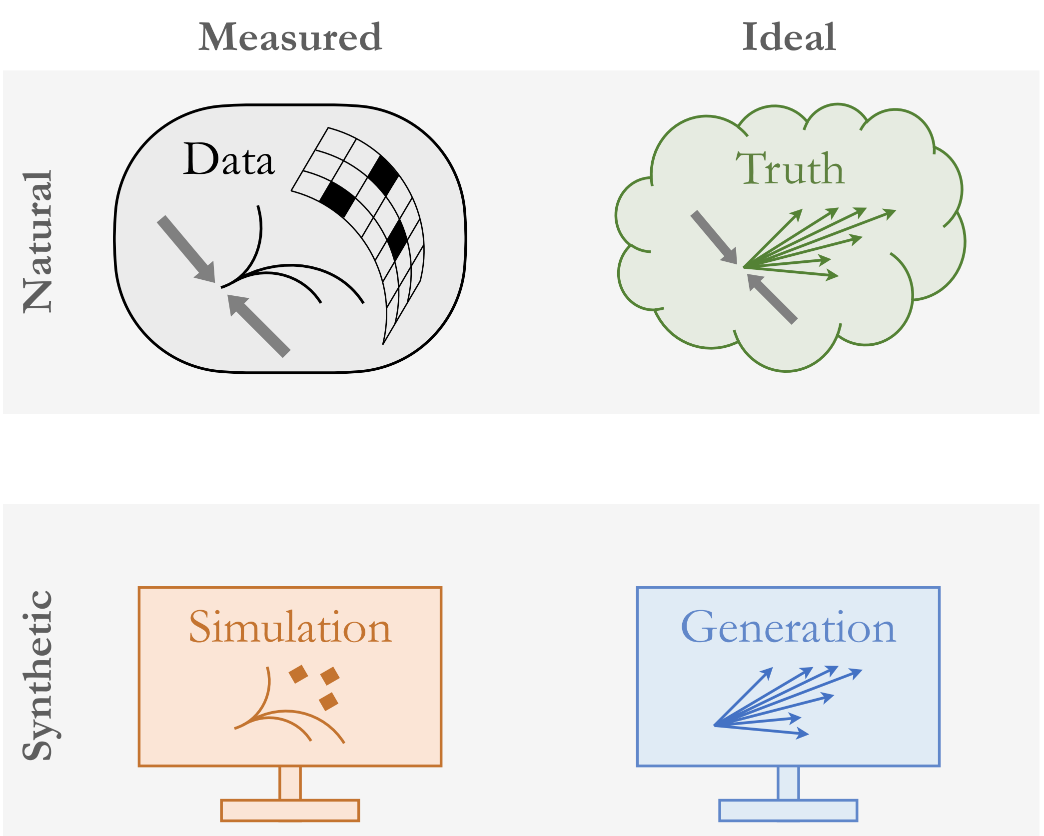
A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001

Unfold by iterating: OmniFold



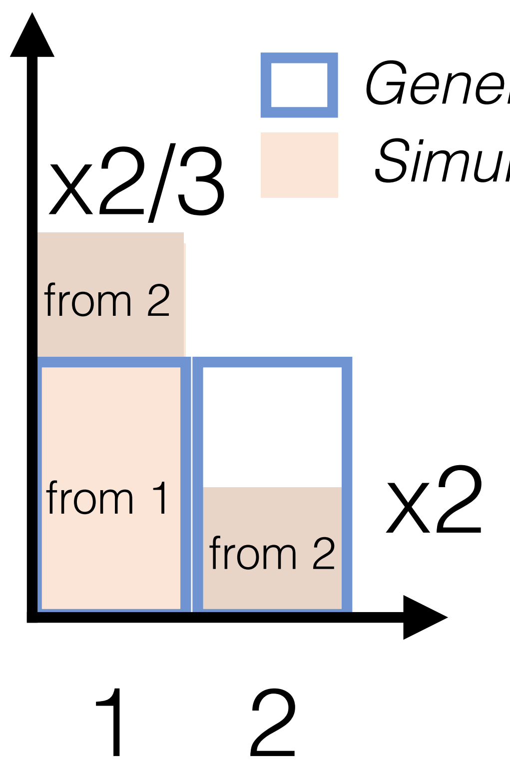


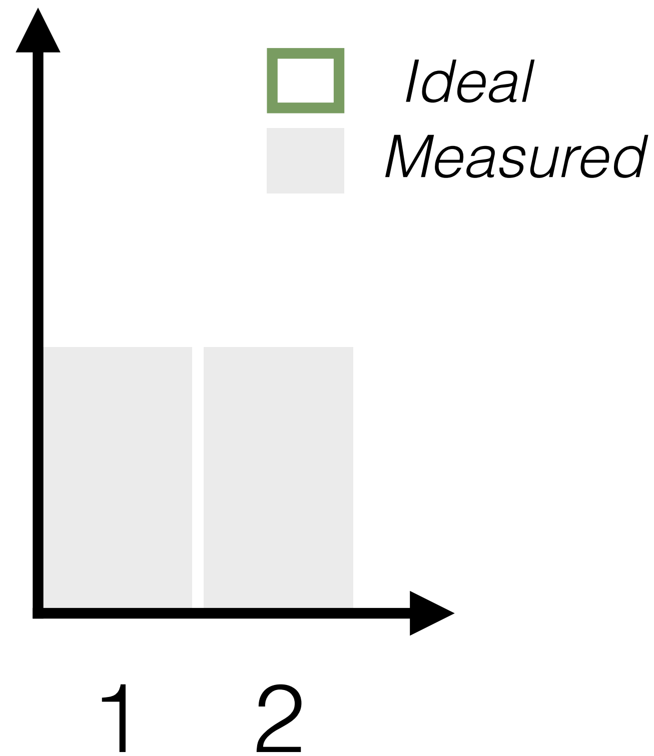
Measured \ Ideal	1	2
2	0%	50%
1	100%	50%



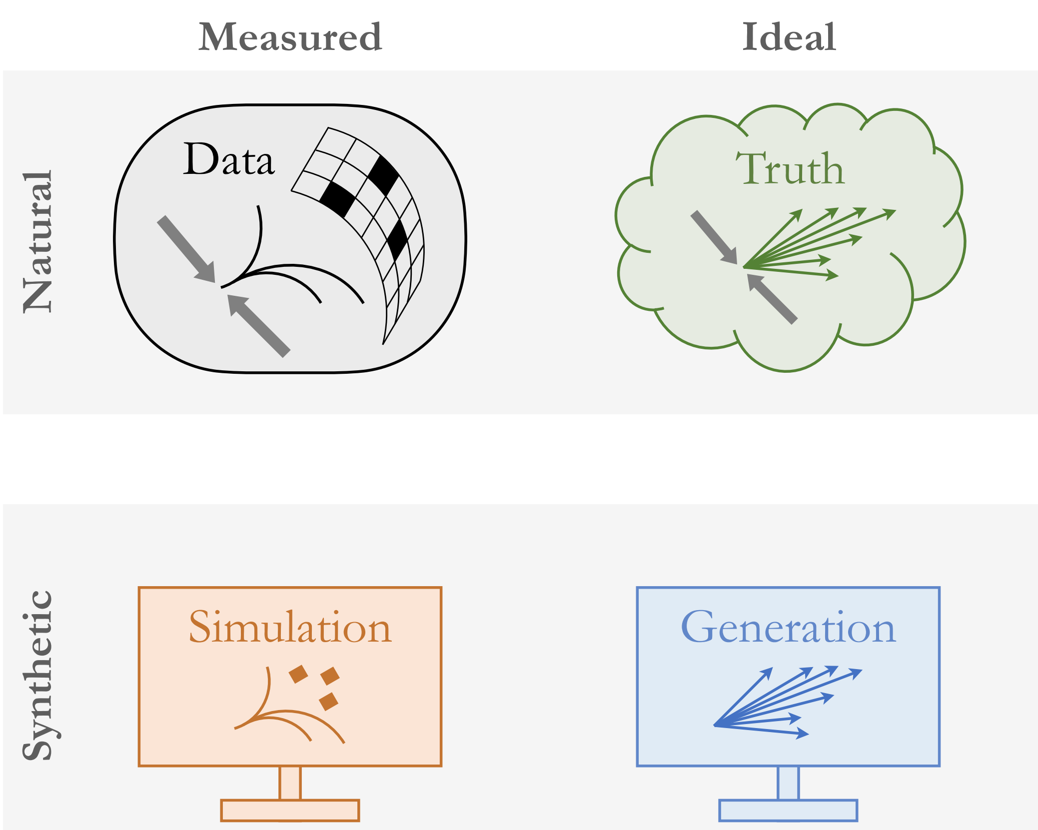
A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001

Unfold by iterating: OmniFold





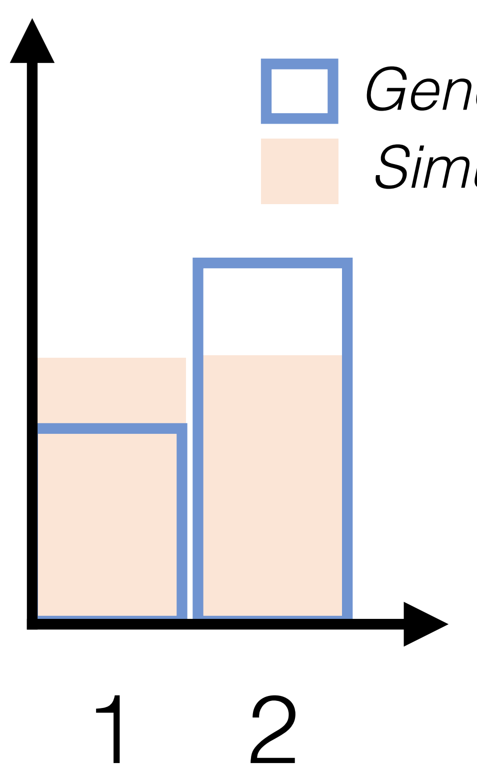
Measured \ Ideal	1	2
2	0%	50%
1	100%	50%

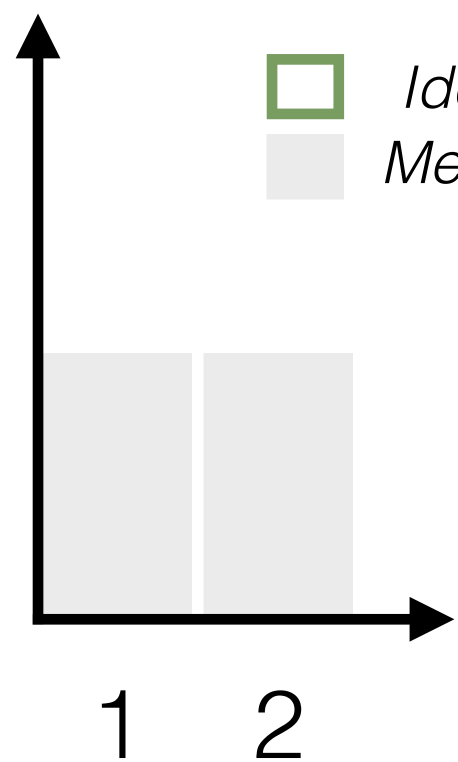


A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001

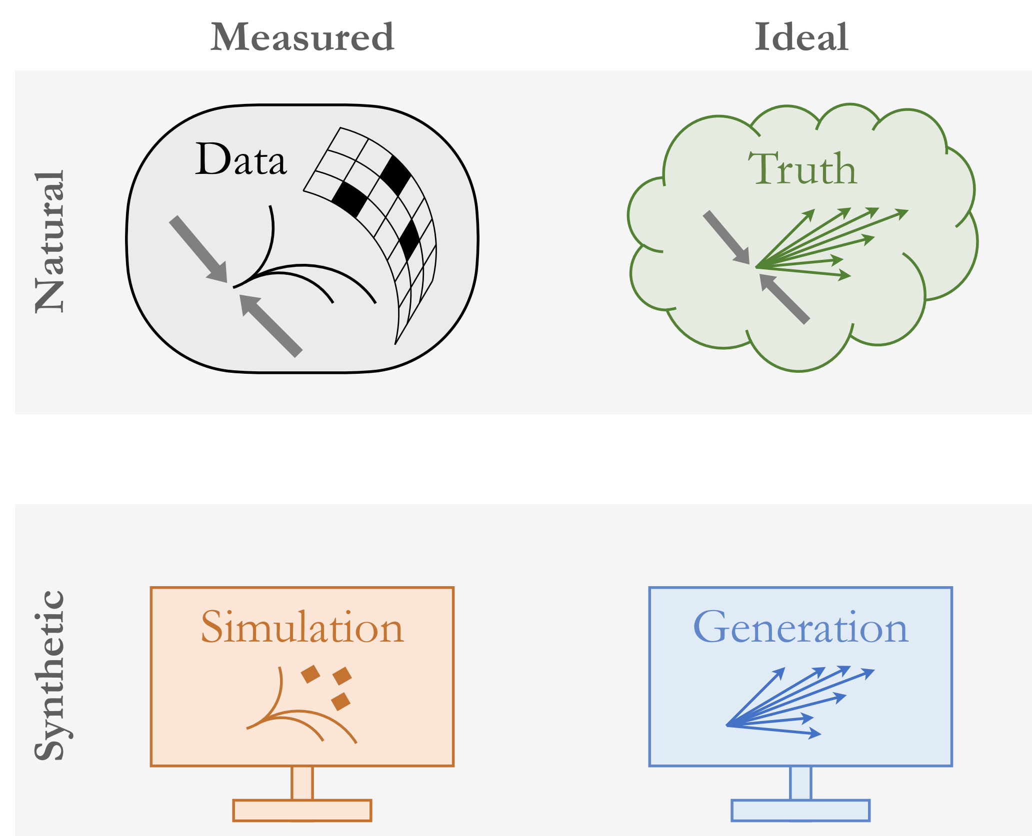
Unfold by iterating: OmniFold

After iteration 1





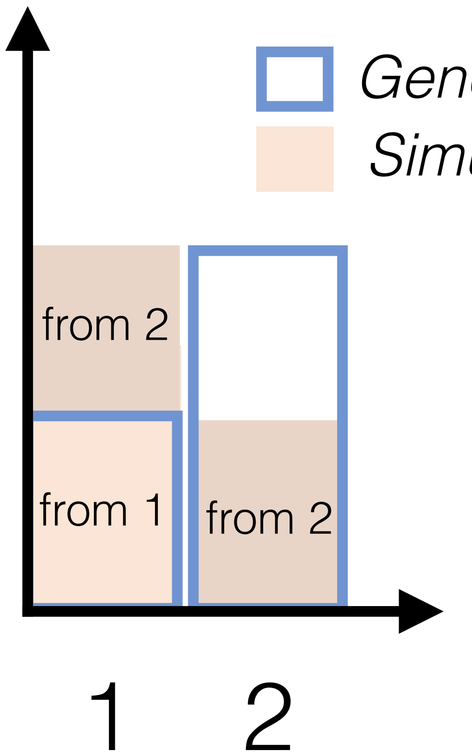
Measured	Ideal: 1	Ideal: 2
2	0%	50%
1	100%	50%

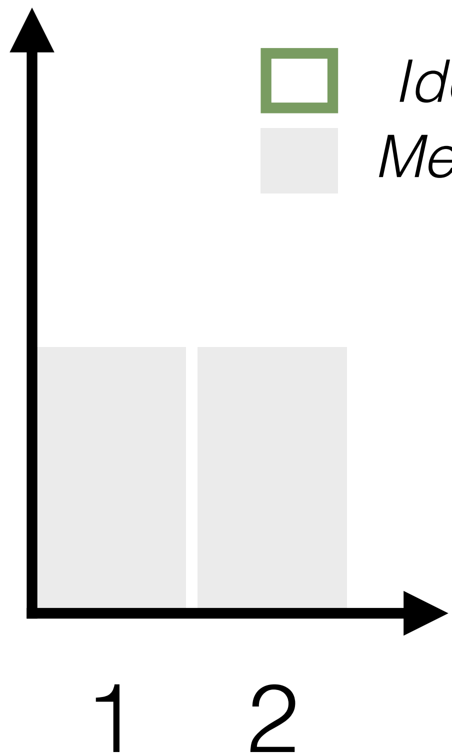


A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001

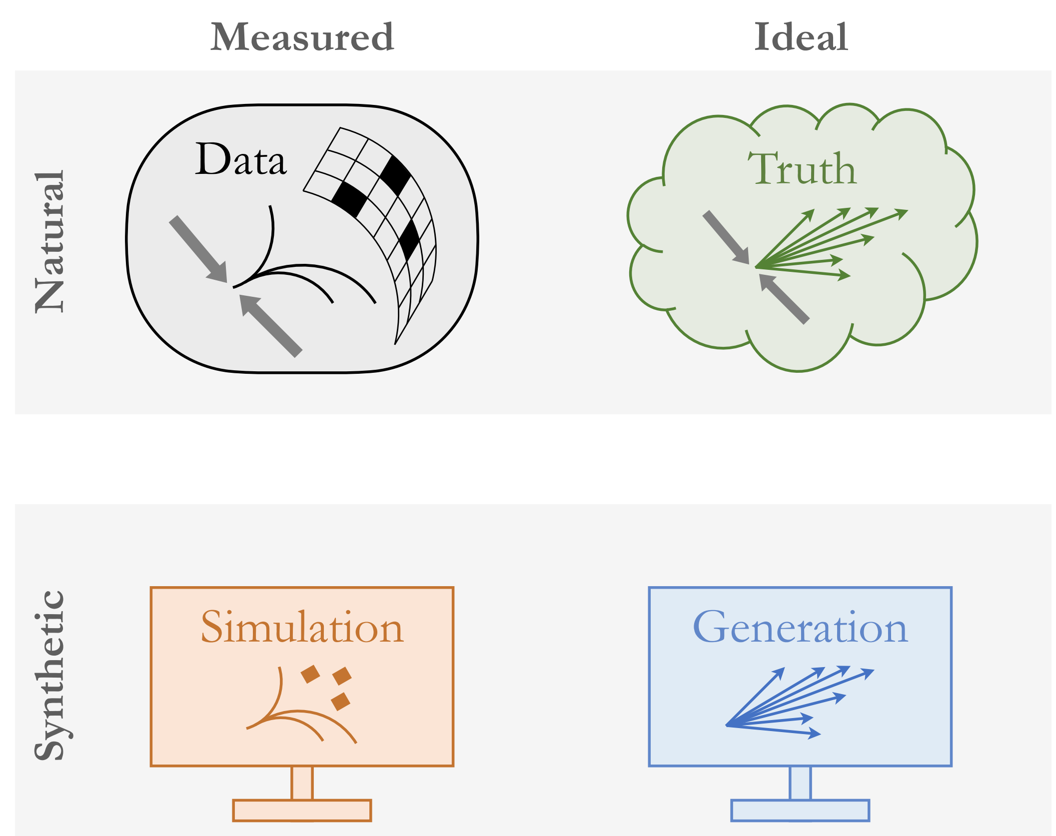
Unfold by iterating: OmniFold

After iteration 1





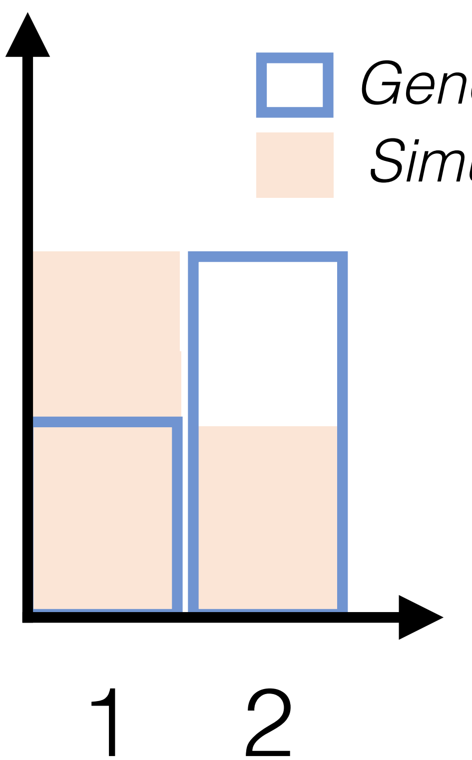
Measured \ Ideal	1	2
2	0%	50%
1	100%	50%

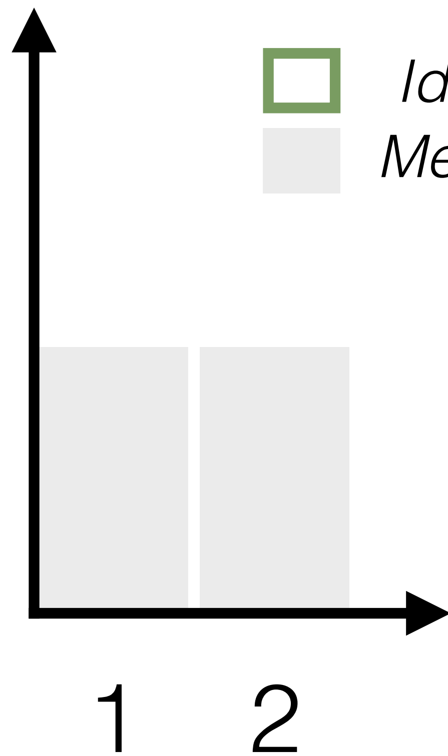


A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001

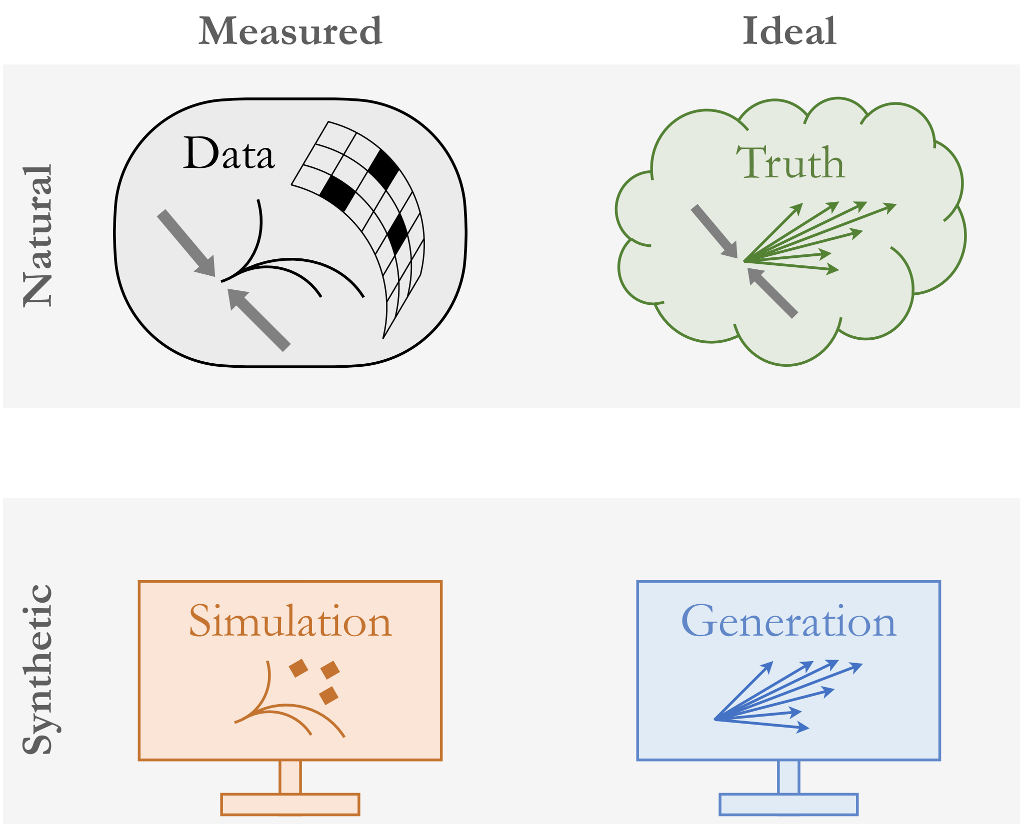
Unfold by iterating: OmniFold

After iteration 1





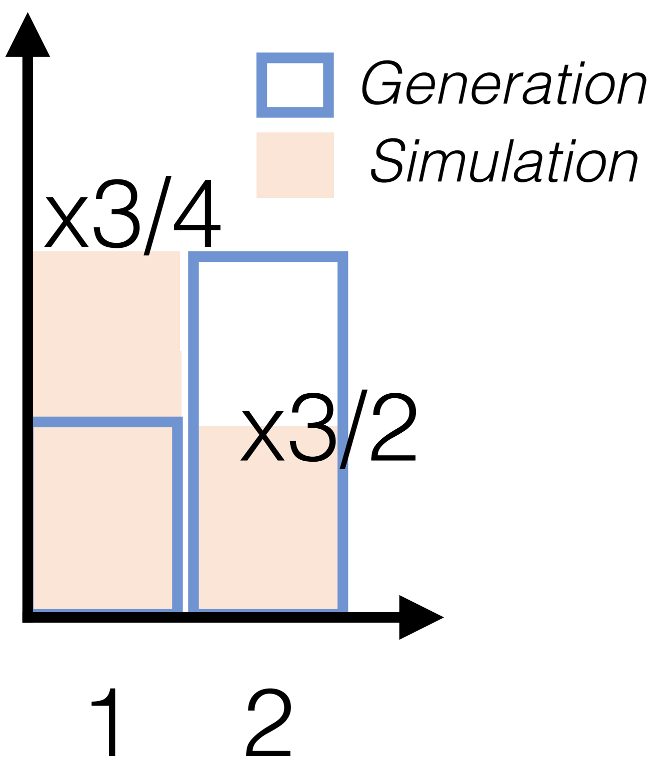
Measured \ Ideal	1	2
2	0%	50%
1	100%	50%

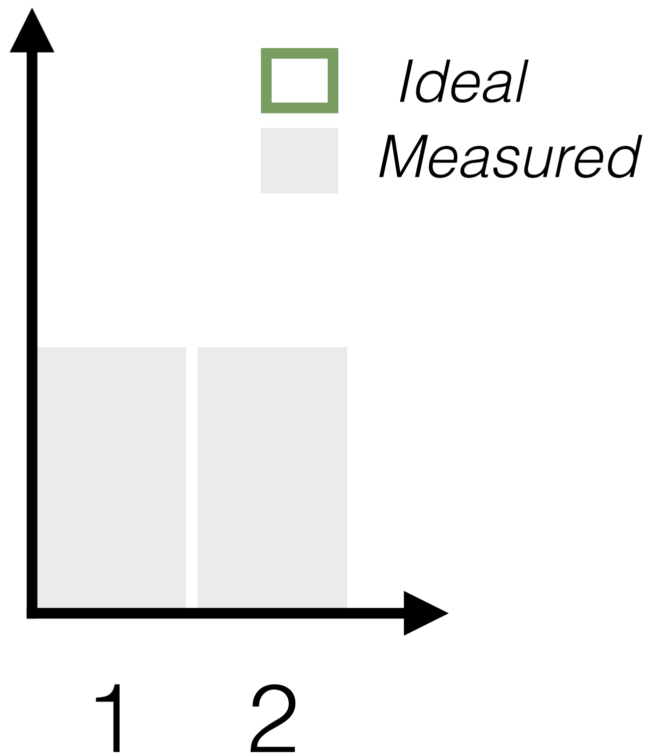


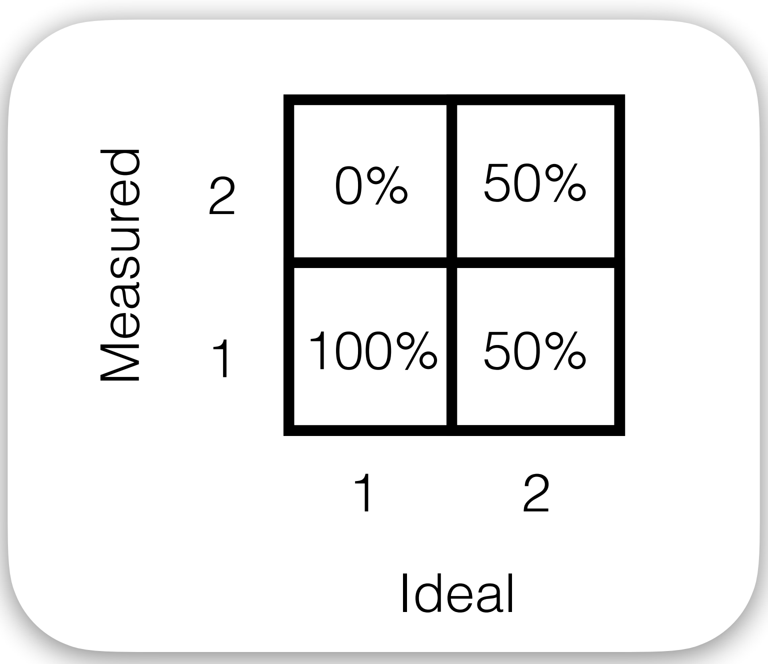
A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001

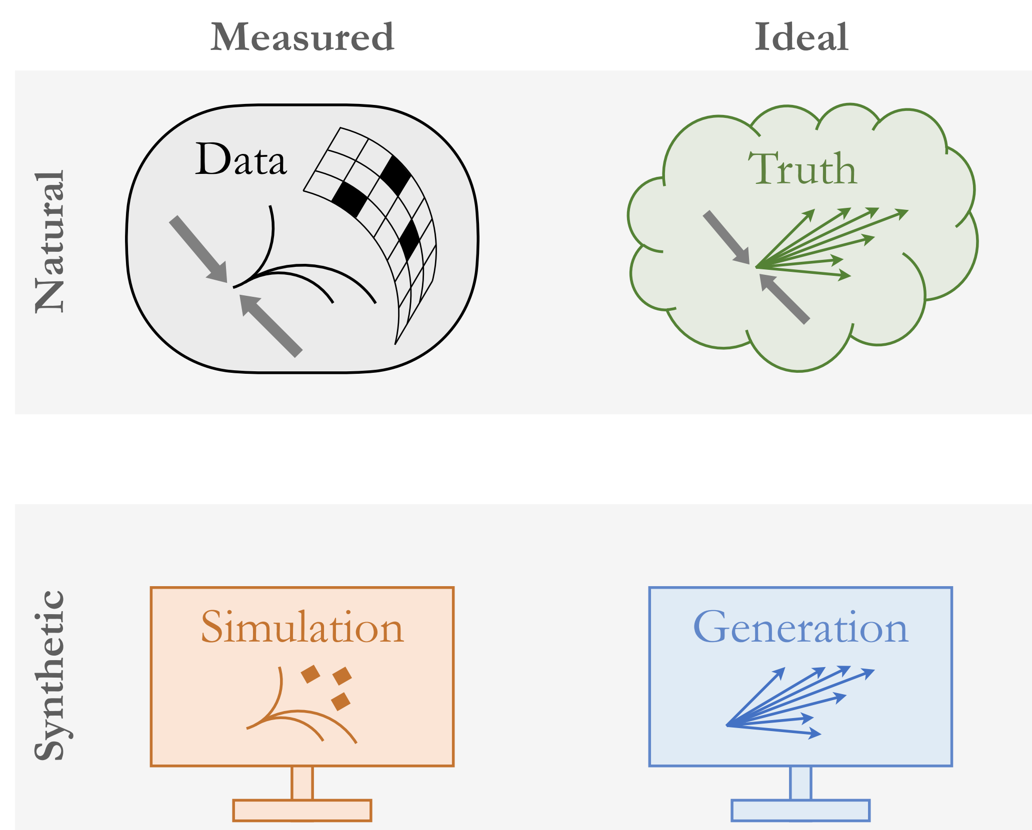
Unfold by iterating: OmniFold

After iteration 1





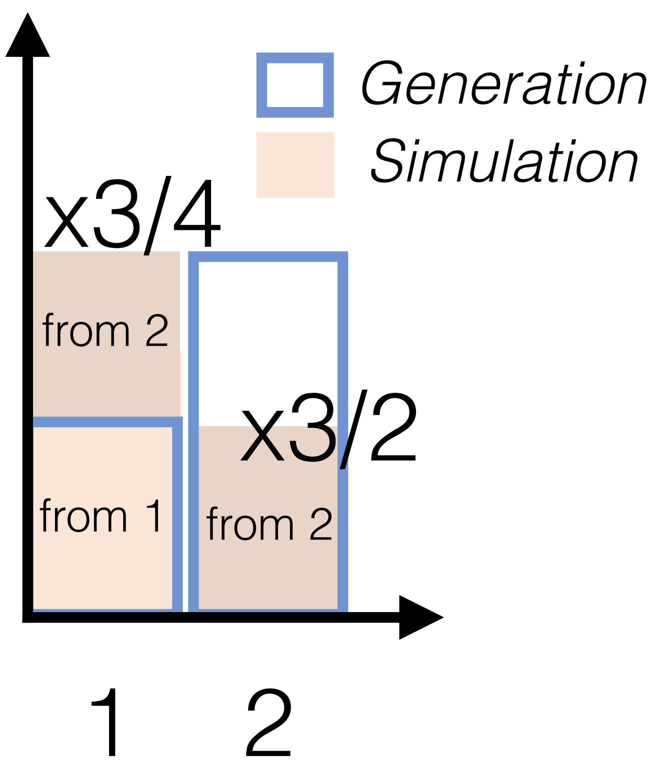


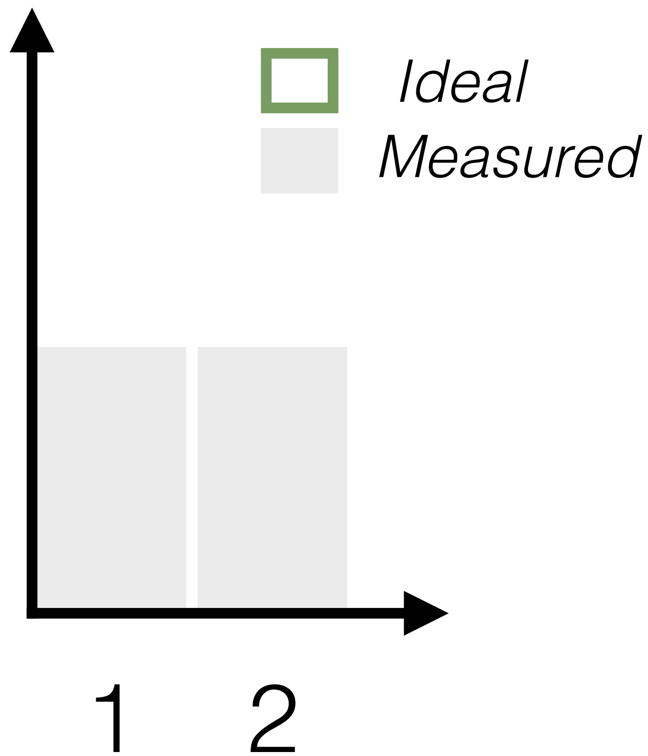


A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001

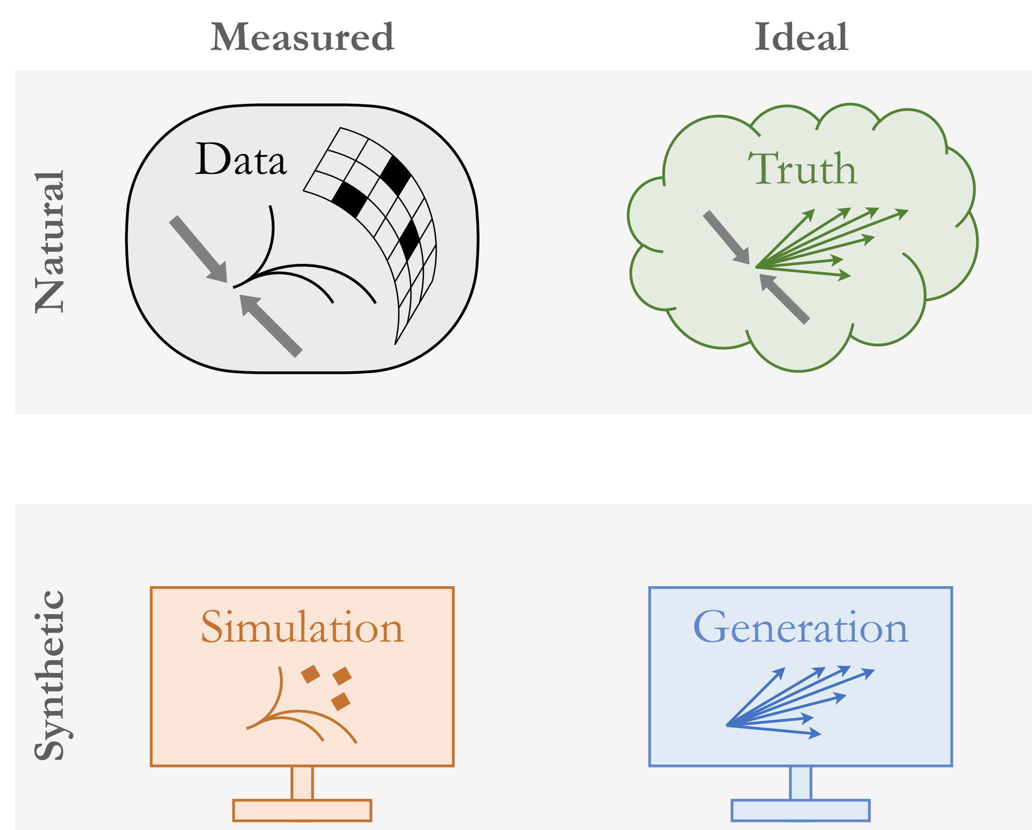
Unfold by iterating: OmniFold

After iteration 1





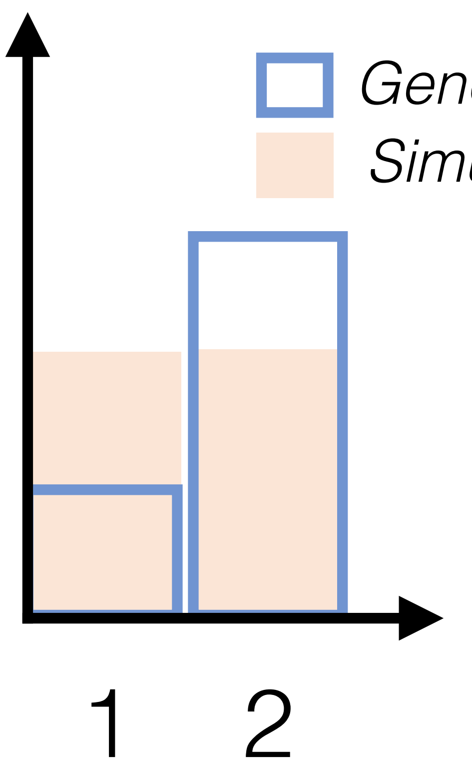
Measured \ Ideal	1	2
2	0%	50%
1	100%	50%

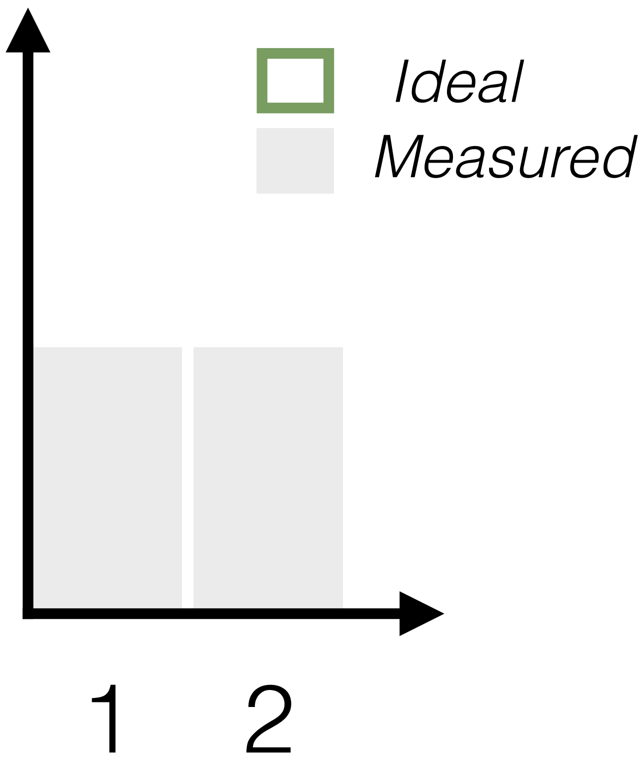


A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001

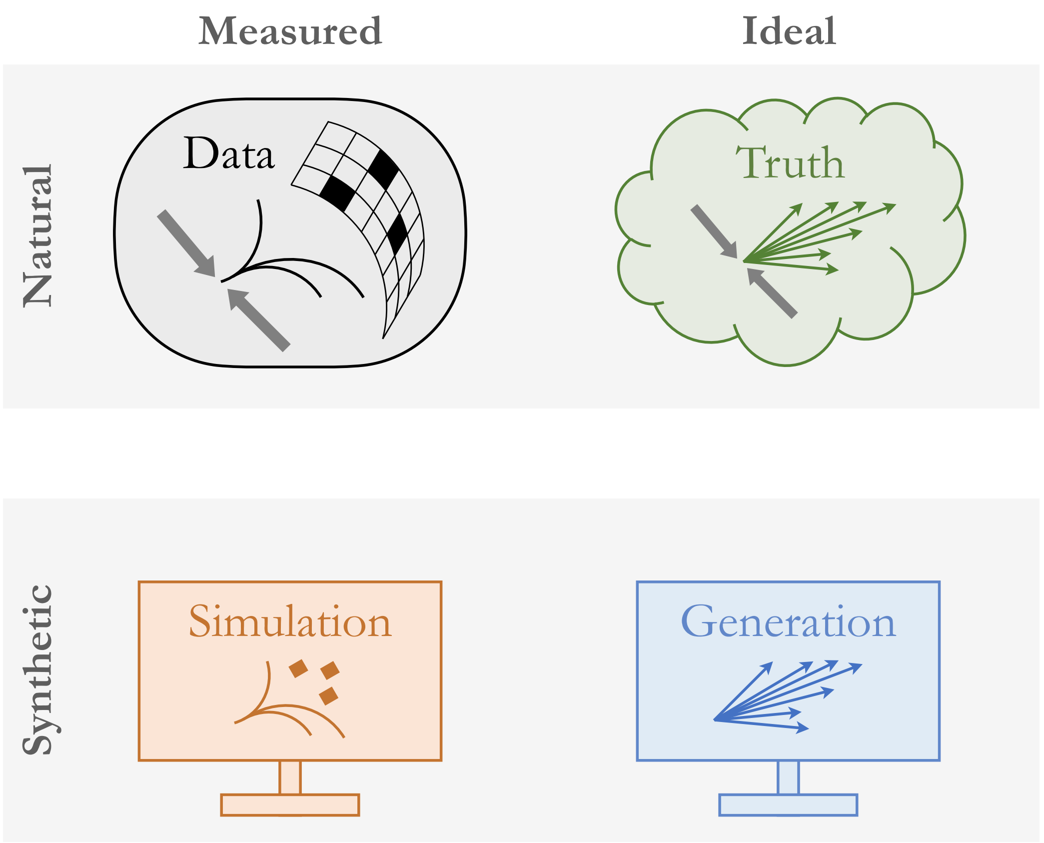
Unfold by iterating: OmniFold

After iteration 2





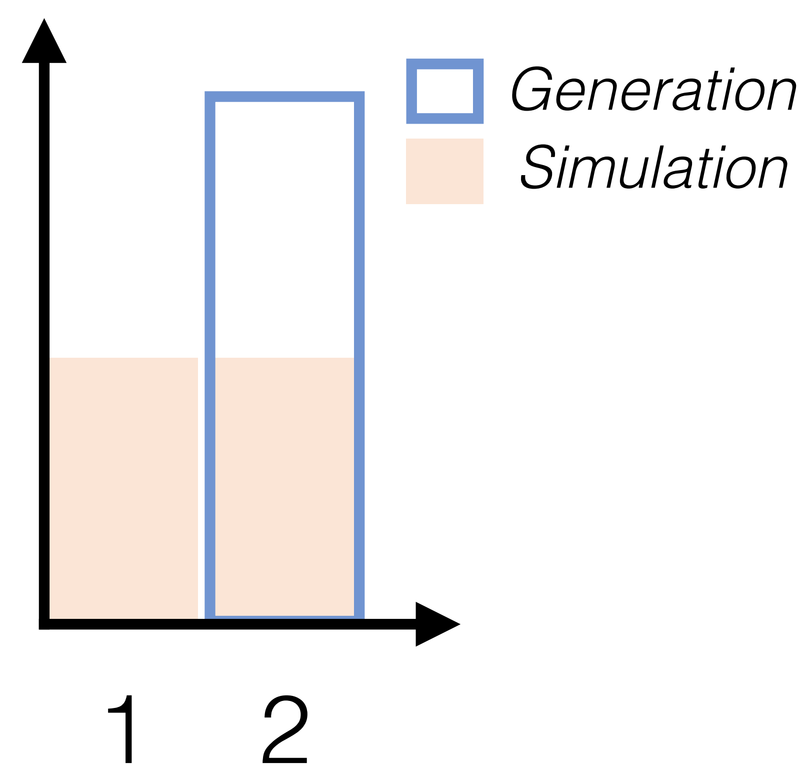
Measured		Ideal 1	Ideal 2
	2	0%	50%
	1	100%	50%

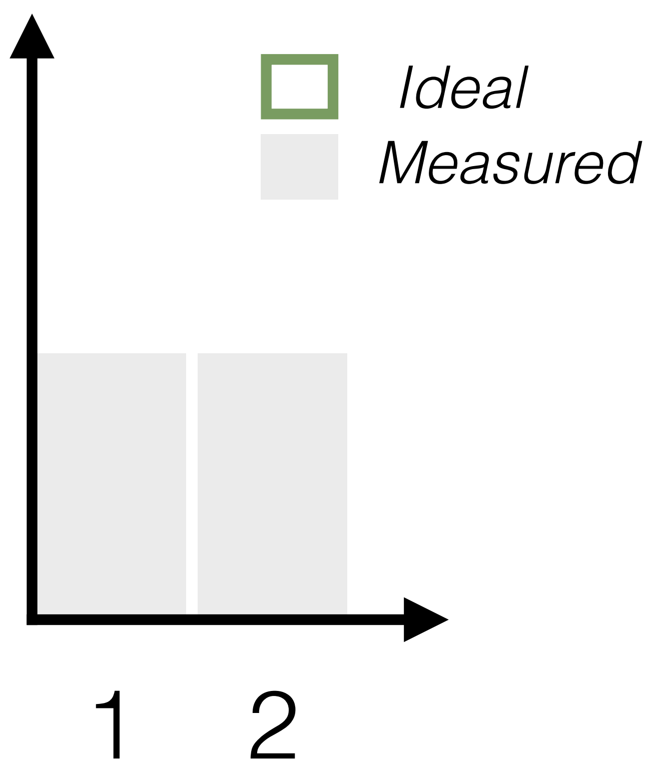


A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001

Unfold by iterating: OmniFold

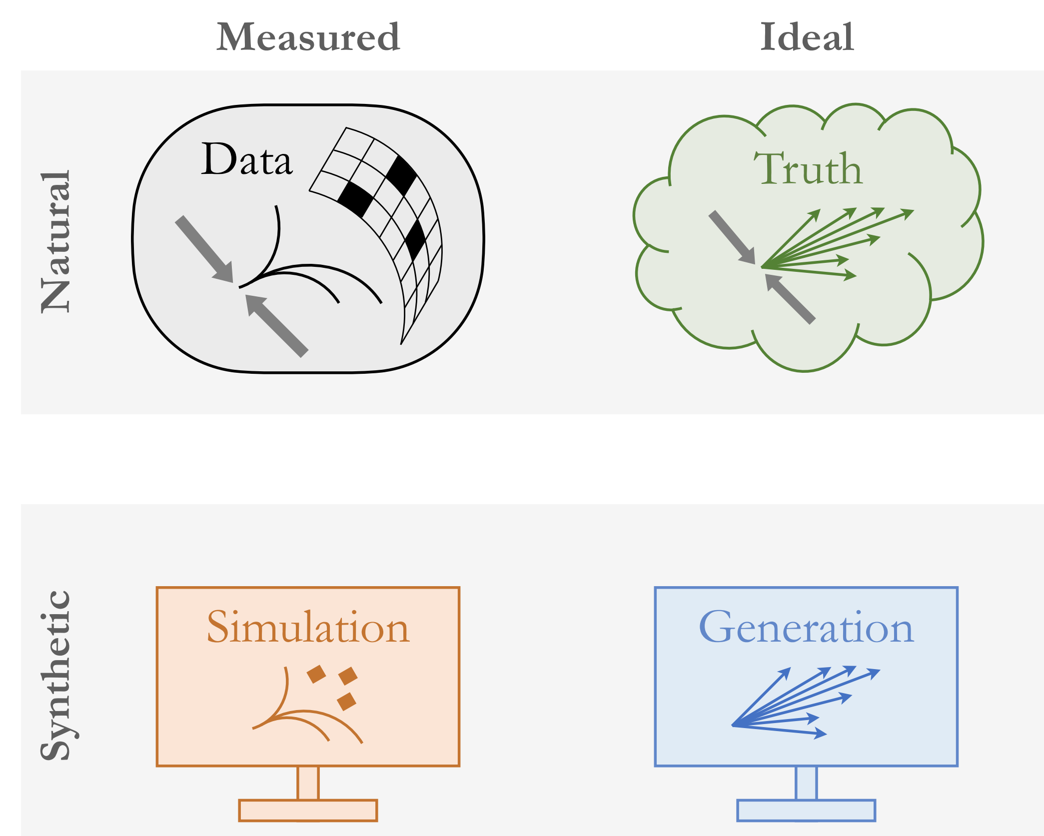
After iteration ∞





N.B. if you just apply p(ideal | measured), *you would have gotten the wrong answer!*

Measured \ Ideal	1	2
2	0%	50%
1	100%	50%



A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001