

Unfolding: Machine Learning Approaches

Benjamin Nachman

Lawrence Berkeley National Laboratory

bpnachman.com

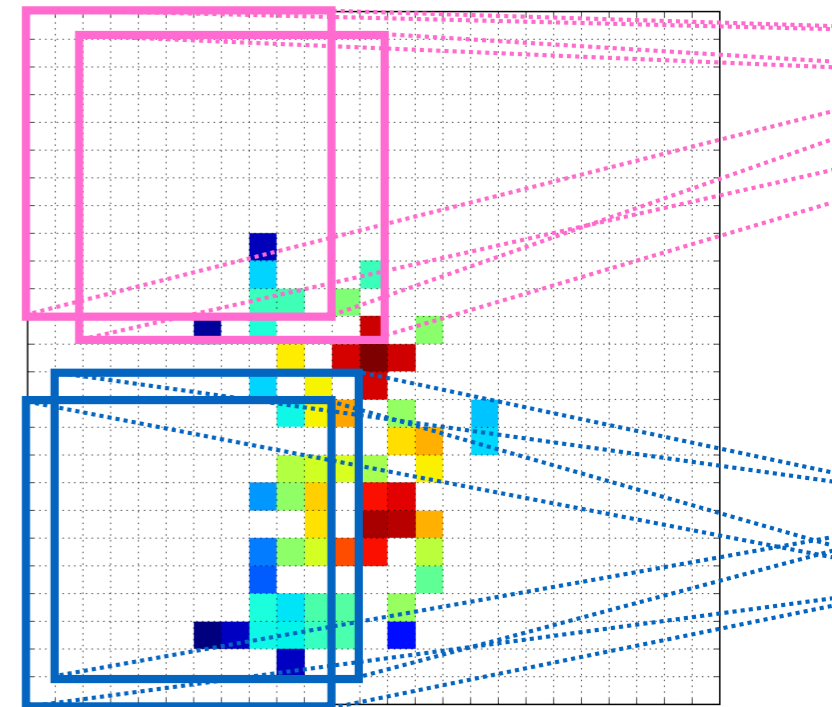
bpnachman@lbl.gov



@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



Deconvolution (“unfolding”):
correcting for detector effects

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

Proton-Proton

Nucleus-Nucleus

Electron-Proton

Neutrino-Nucleus

Cosmic Rays

Electron-Positron

Particle/Nuclear/Astro Physics Experiments

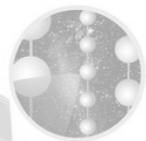
Unfolding

5

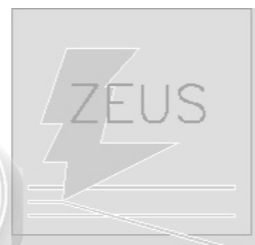
Proton-Proton



Nucleus-Nucleus



ICECUBE
SOUTH POLE NEUTRINO OBSERVATORY



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”?

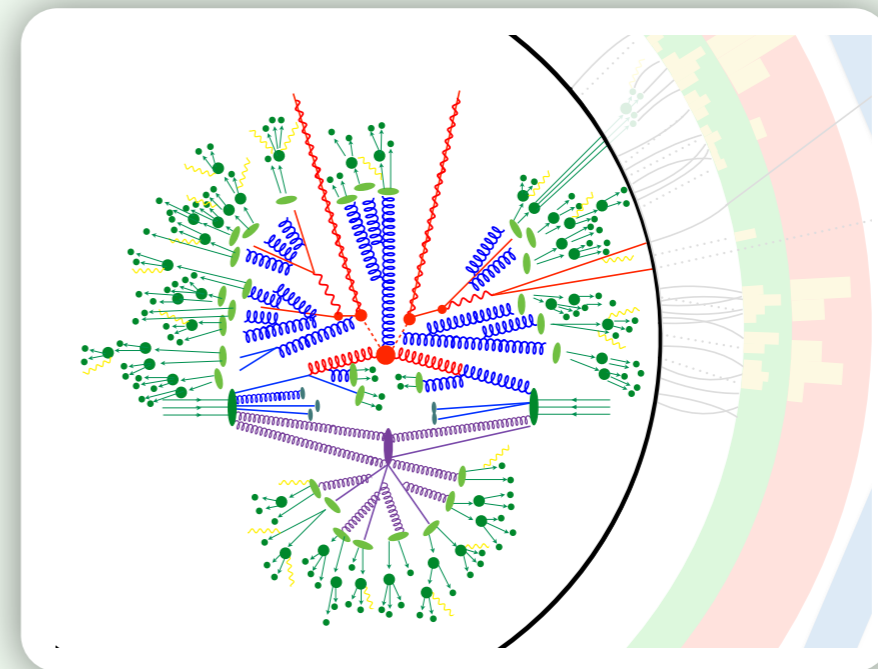
Unfolding is ill-posed, BUT only way to compare different experiments and to compare with non fully exclusive predictions. Data also survive much longer.

The Unfolding Challenge

The Unfolding Challenge

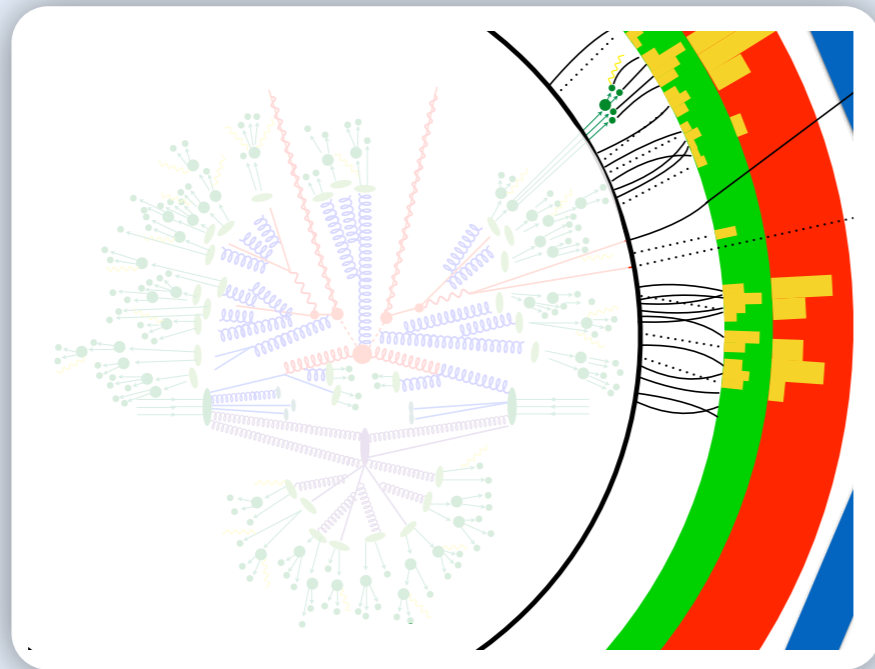
Particle
Level

Want this

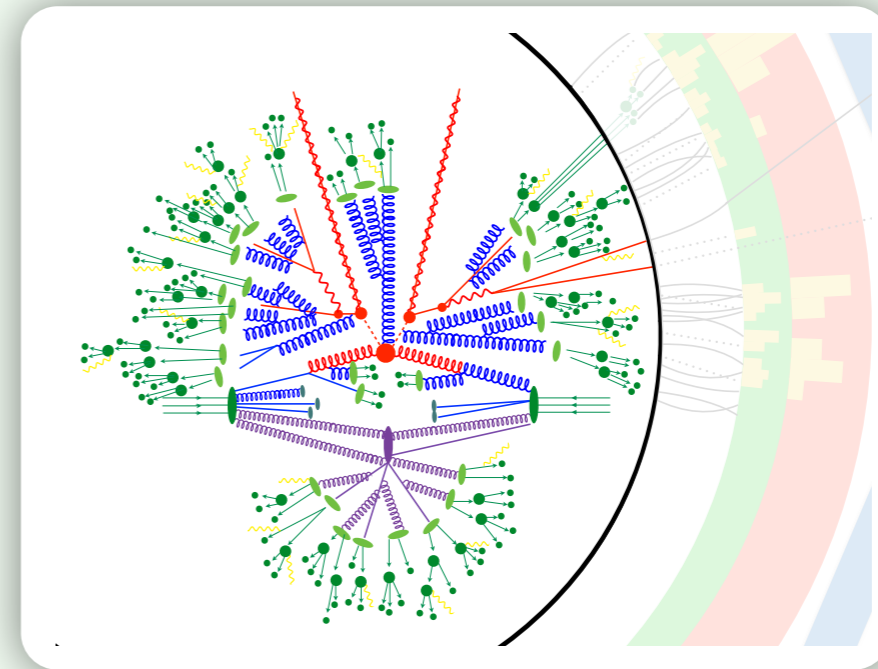


The Unfolding Challenge

Measure this

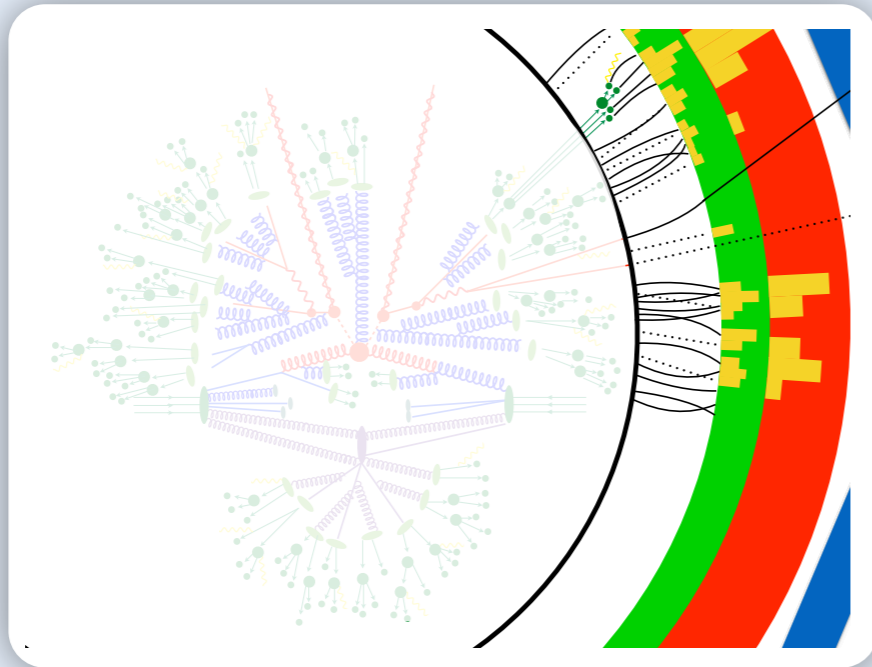


Want this

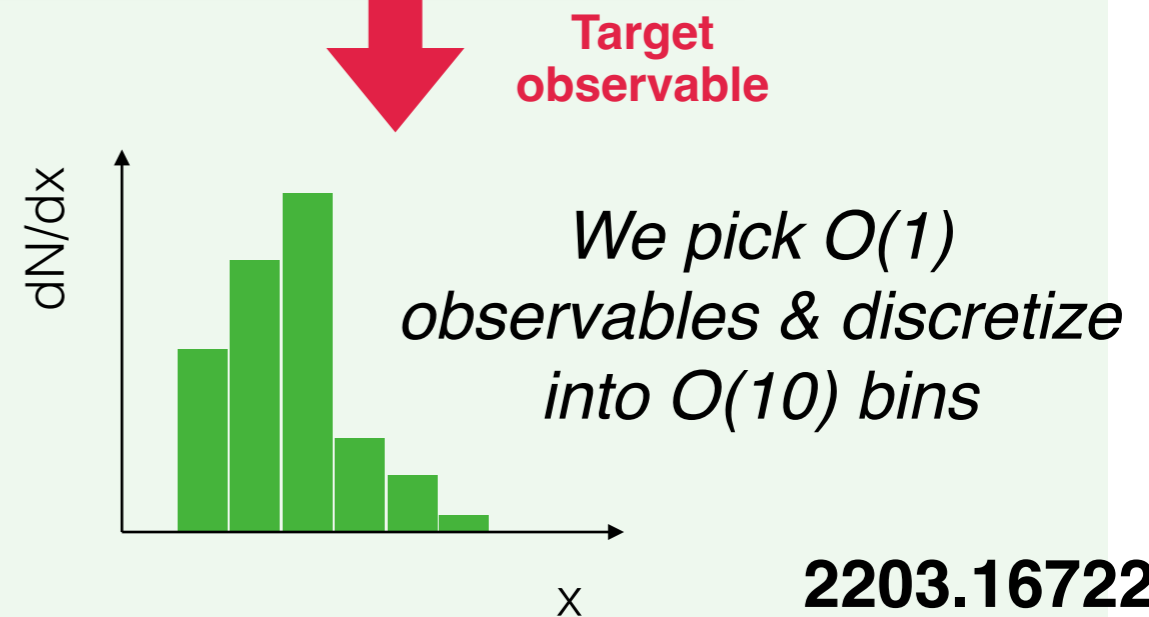
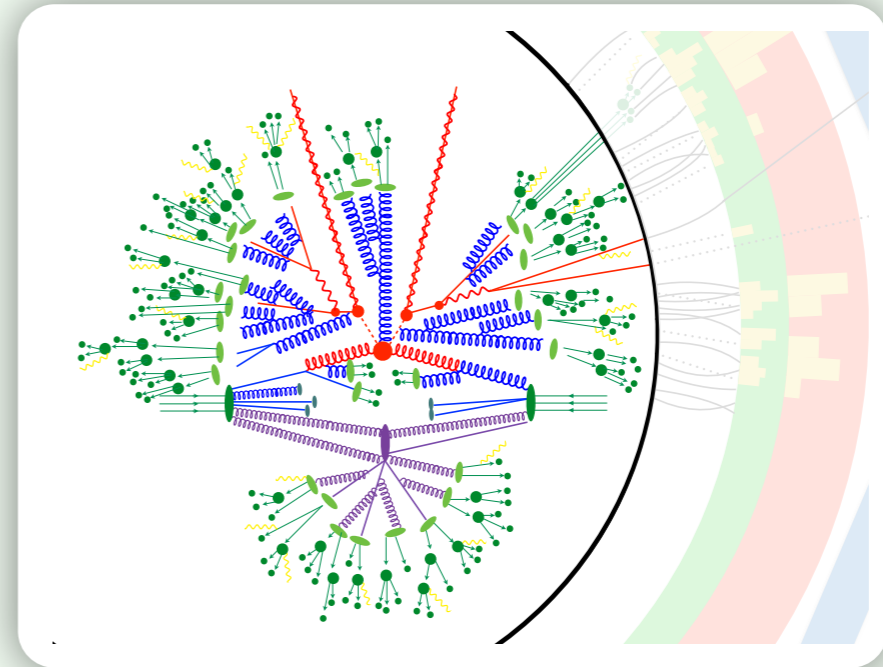


The Unfolding Challenge

Measure this

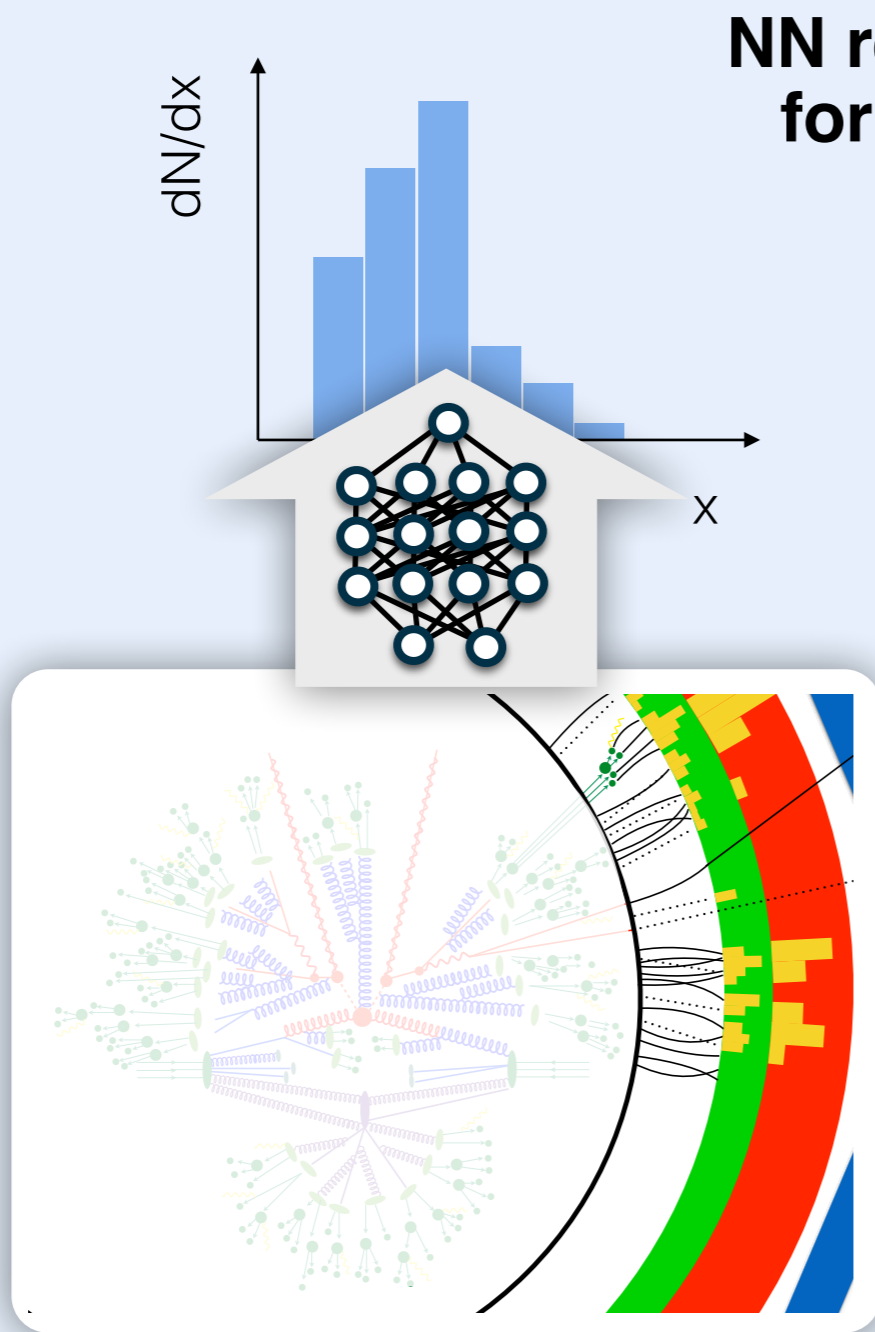


Want this

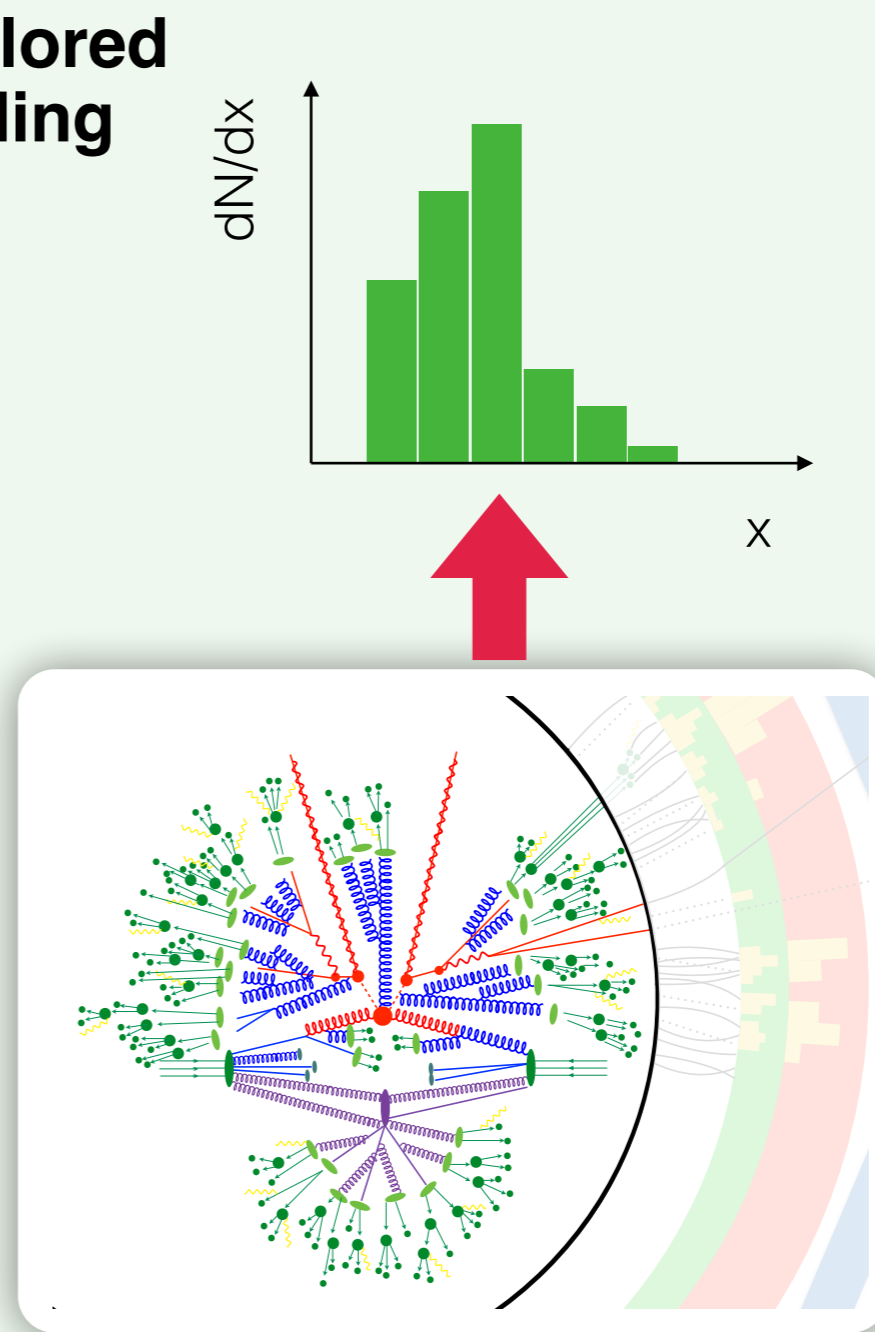


Detector Level

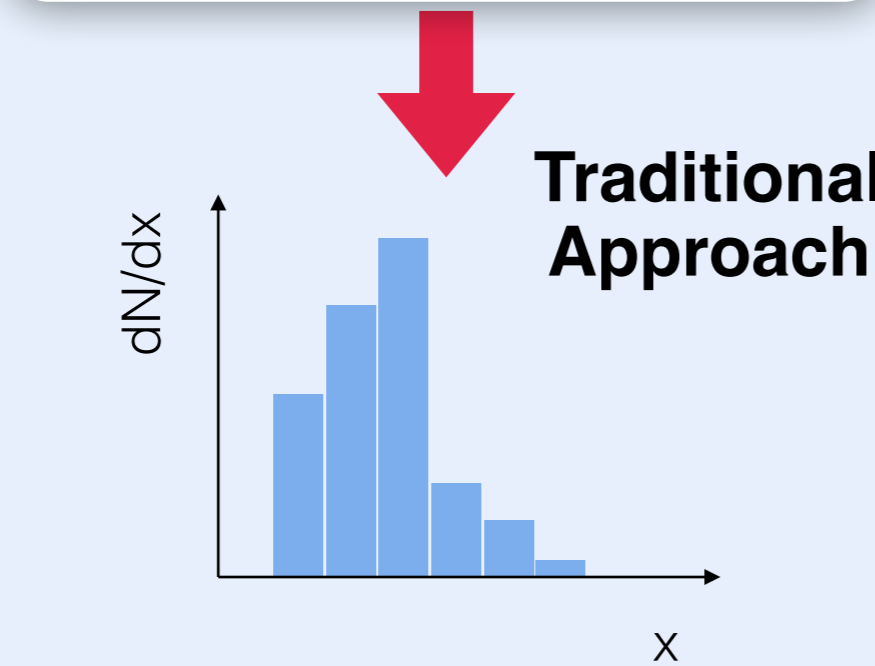
M. Arratia, D. Britzger, O. Long,
BPN, JINST 17 (2022) P07009
(see also A. Glazov, 1712.01814)



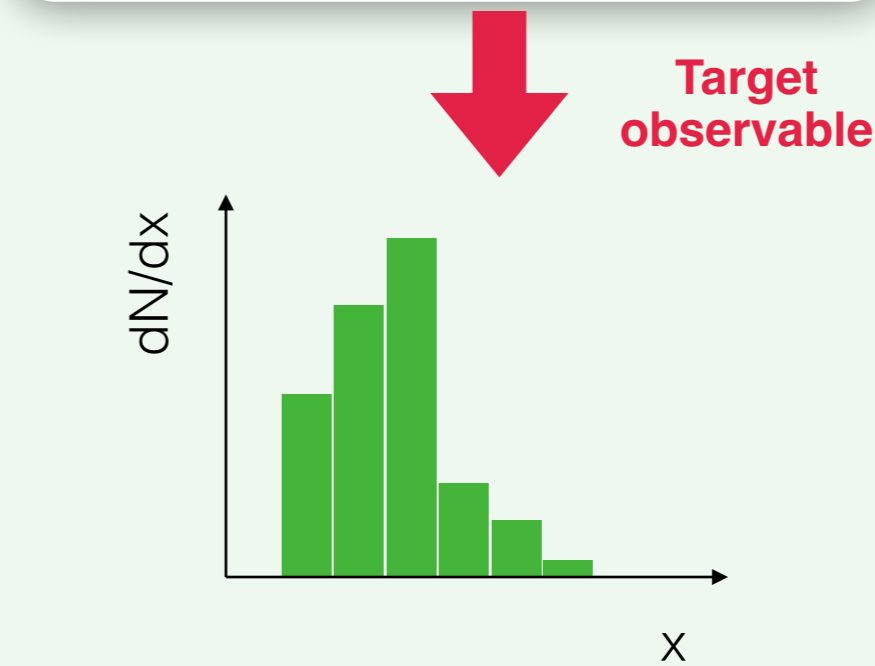
NN reco tailored for unfolding



Particle Level

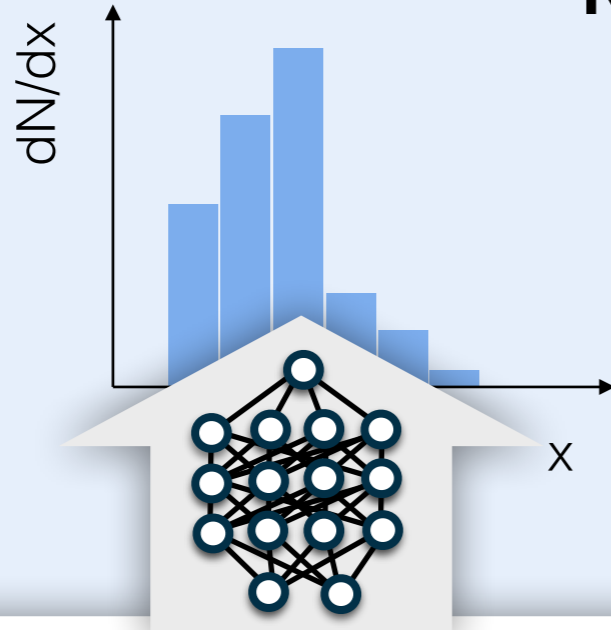


Traditional Approach

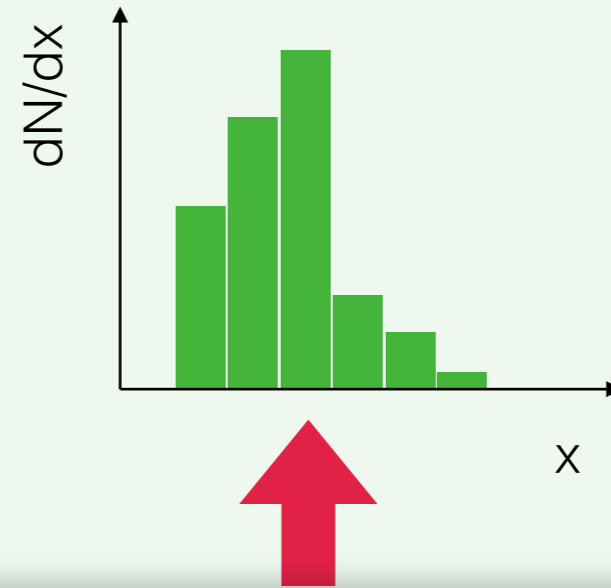
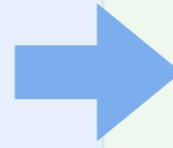


Target observable

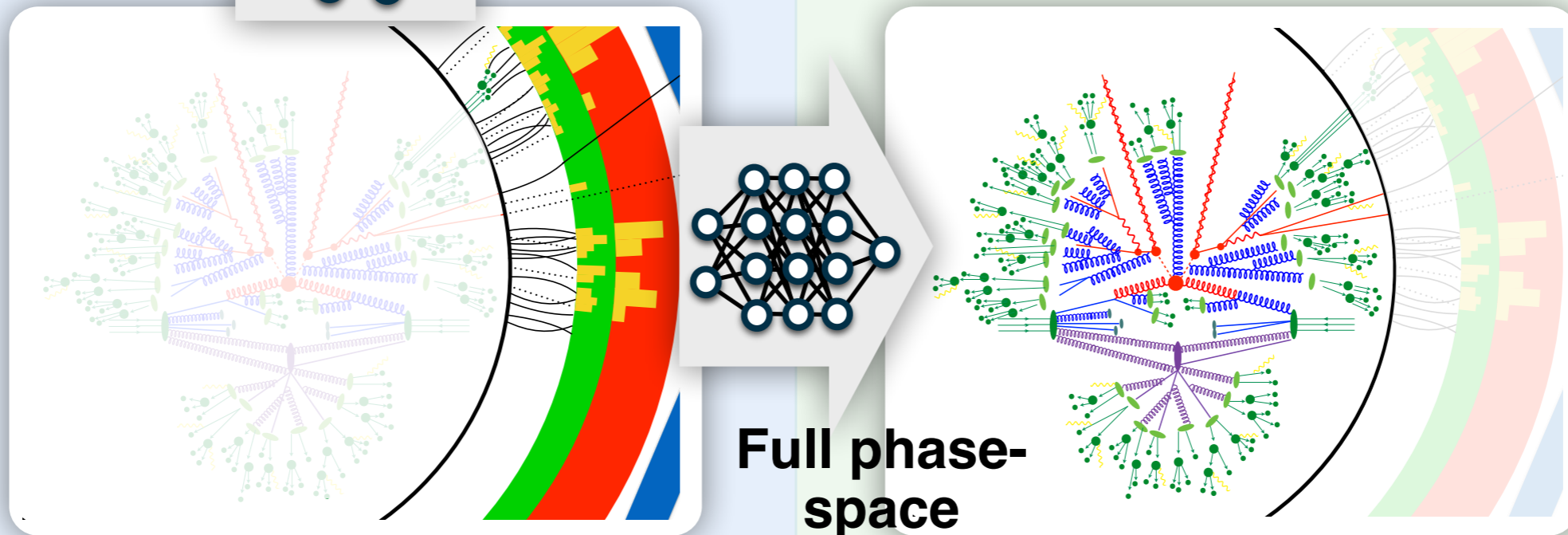
Detector Level



NN reco tailored for unfolding



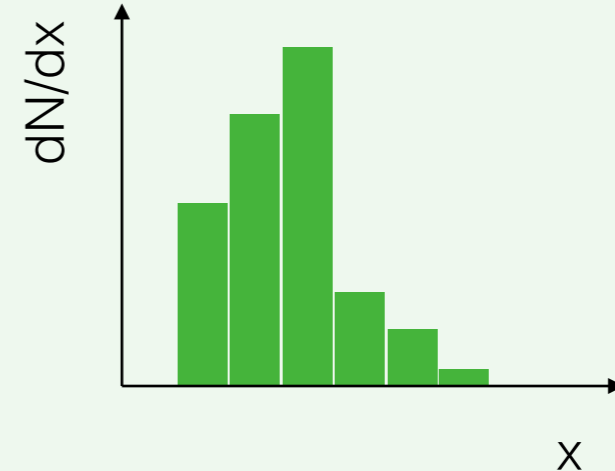
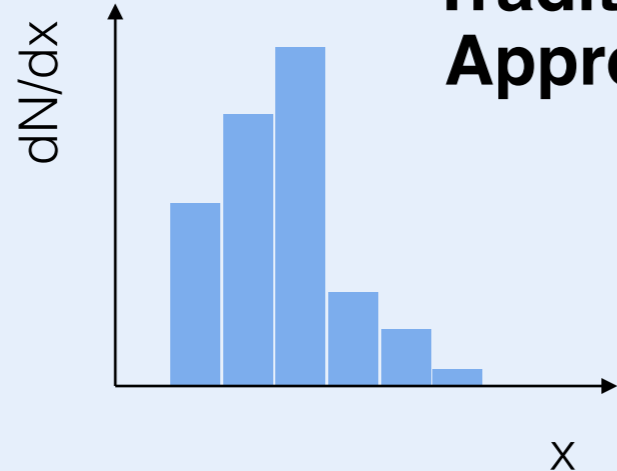
Particle Level



Full phase-space unfolding

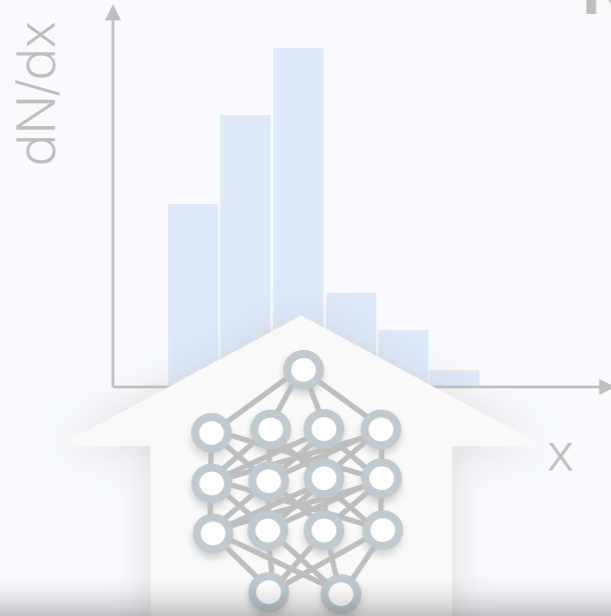


Traditional Approach

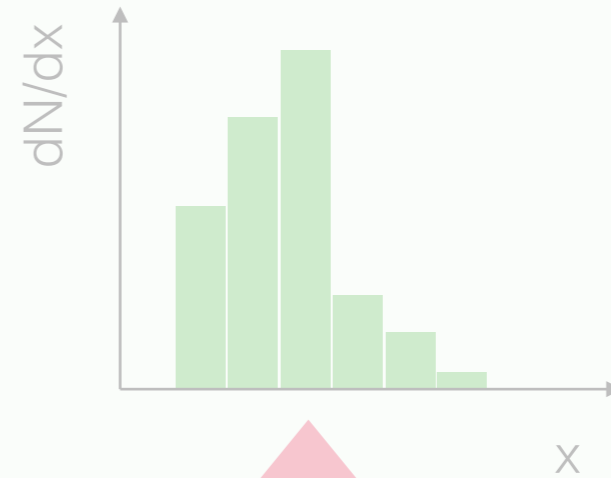
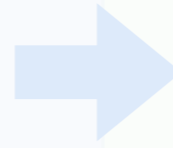


Target observable

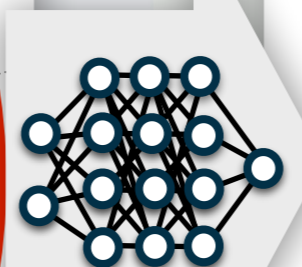
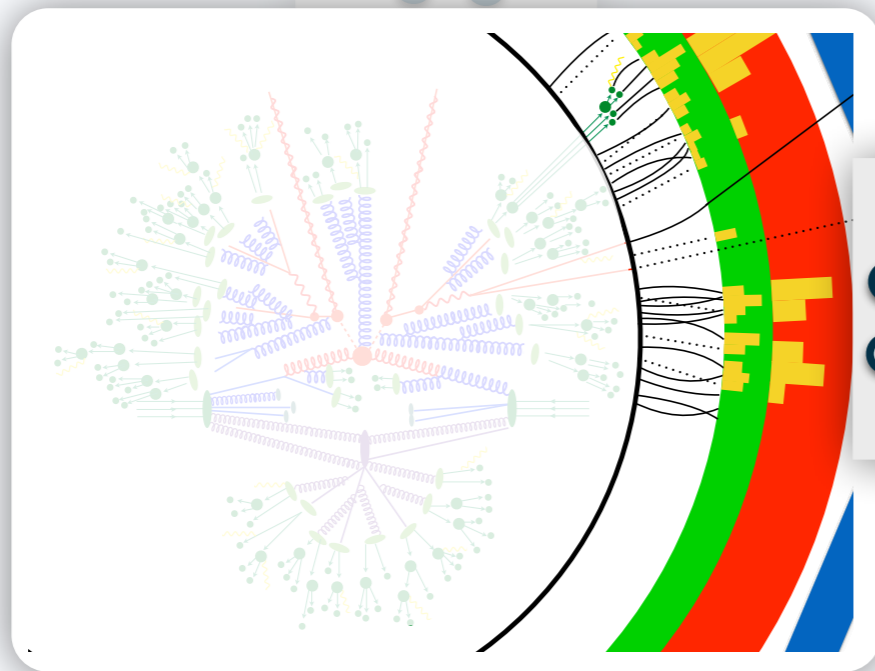
Detector Level



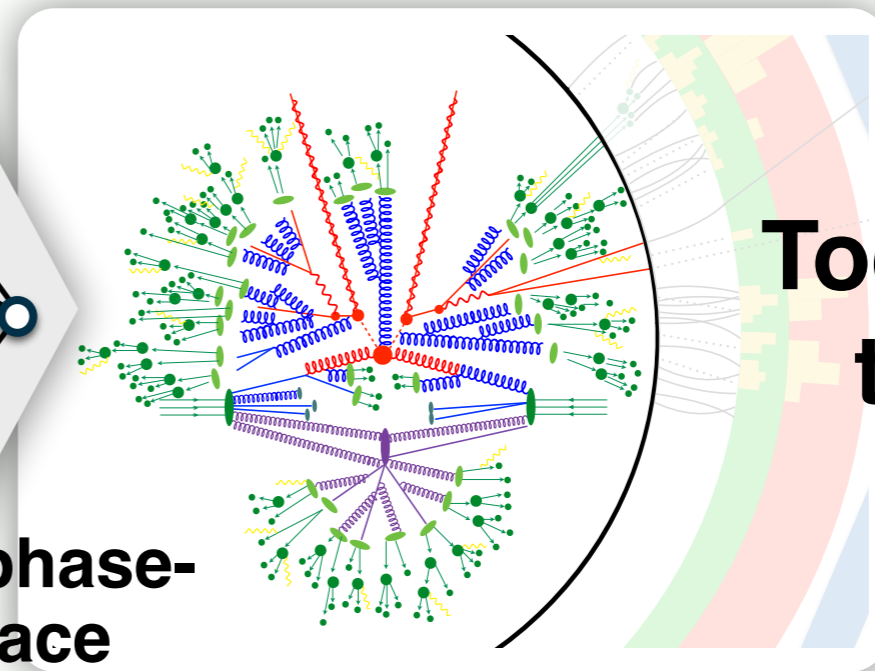
NN reco tailored for unfolding



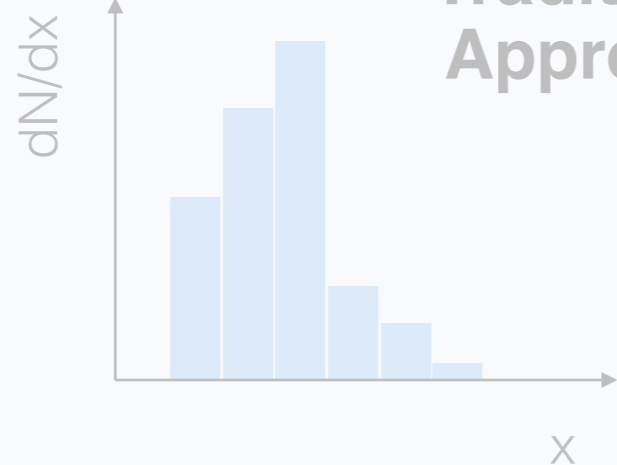
Particle Level



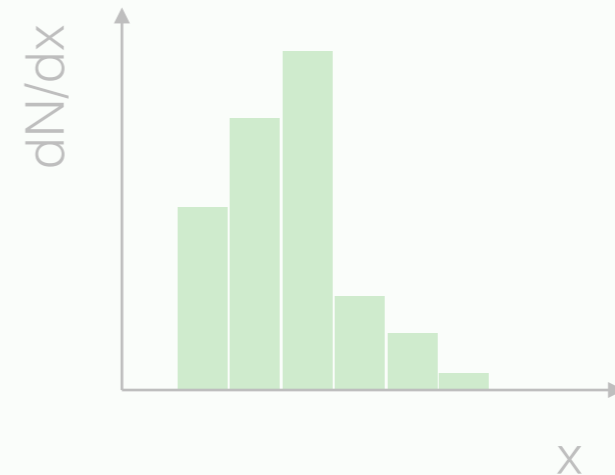
Full phase-space unfolding



Today's talk



Traditional Approach



Target observable

Why unbinned (+high-dimensional)?

13



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

Why unbinned (+high-dimensional)?

14

Inference-Aware Binning

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

What about moments?

(see also K. Desai, BPN, J. Thaler, [\[paper\]](#))

Why unbinned (+high-dimensional)?

15

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 reuse results for a function of the phase space.

Why unbinned (+high-dimensional)?

16

Inference-Aware Binning

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

What about moments?

(see also K. Desai, BPN, J. Thaler, [\[paper\]](#))

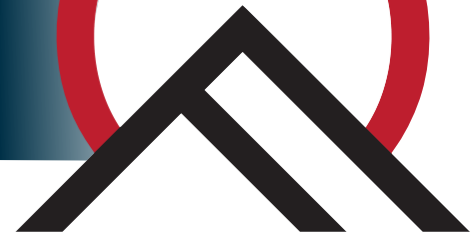
Derivative Measurements

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?



Classifier-Based Methods

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

Classifier-Based Methods

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

Density-Based Methods

*Learn (unfolded) data probably
density implicitly or explicitly.*

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.*

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. Diefenbacher, G. Liu, V. Mikuni, B. Nachman, W. Nie, 2308.12351

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. Diefenbacher, G. Liu, V. Mikuni, B. Nachman, W. Nie, 2308.12351*

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)*



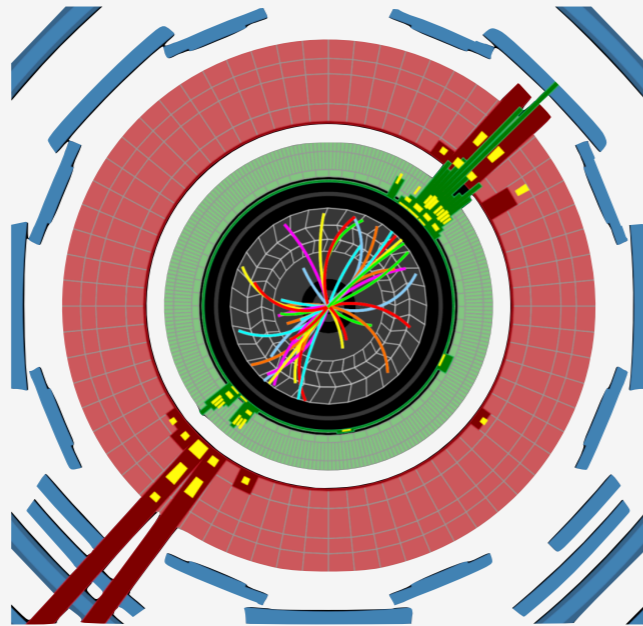
My focus will be on a method called **OmniFold**.

A brief introduction to OmniFold

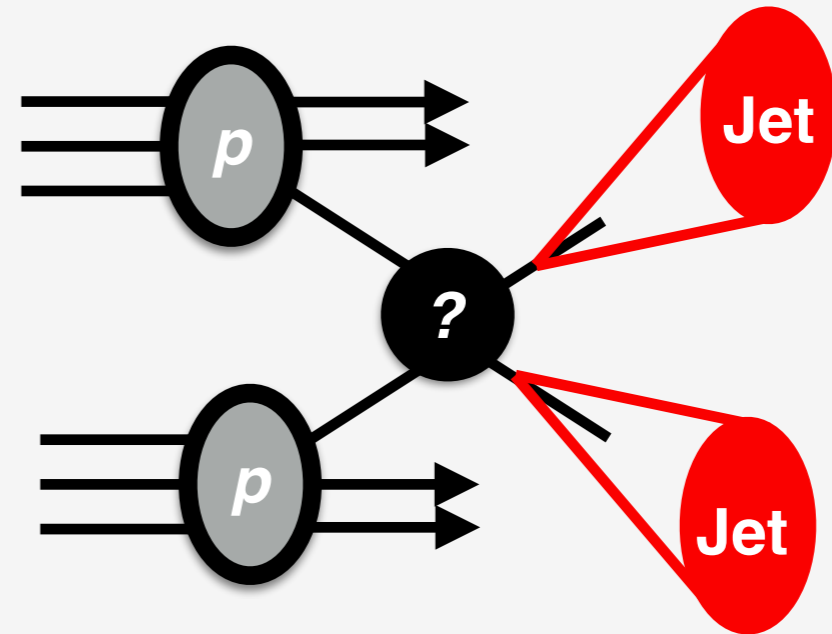
23

Nature

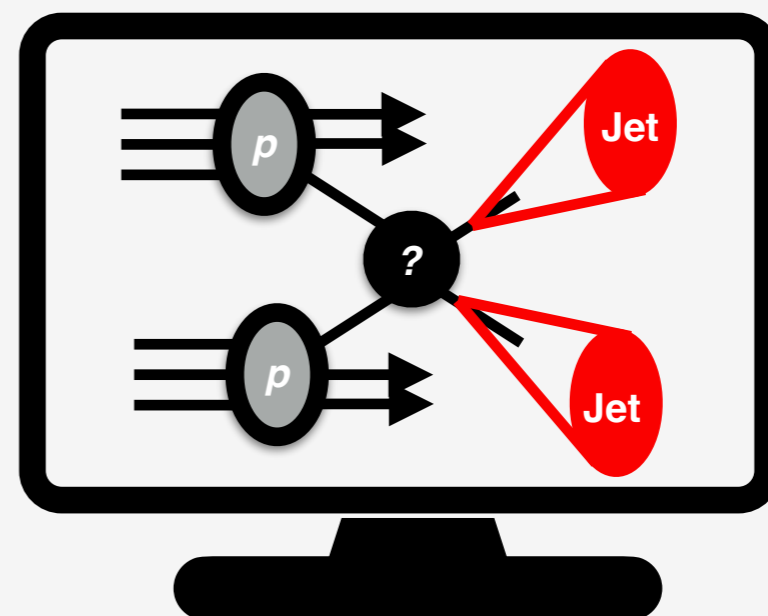
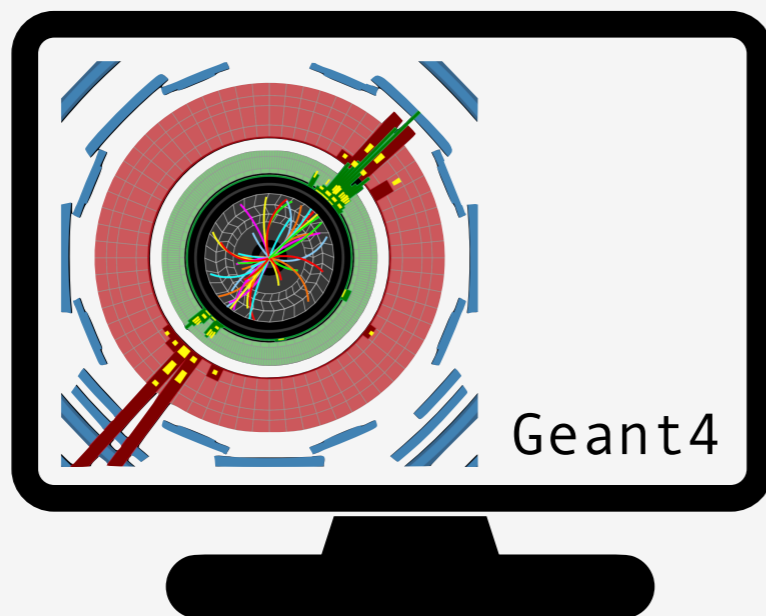
Detector-level



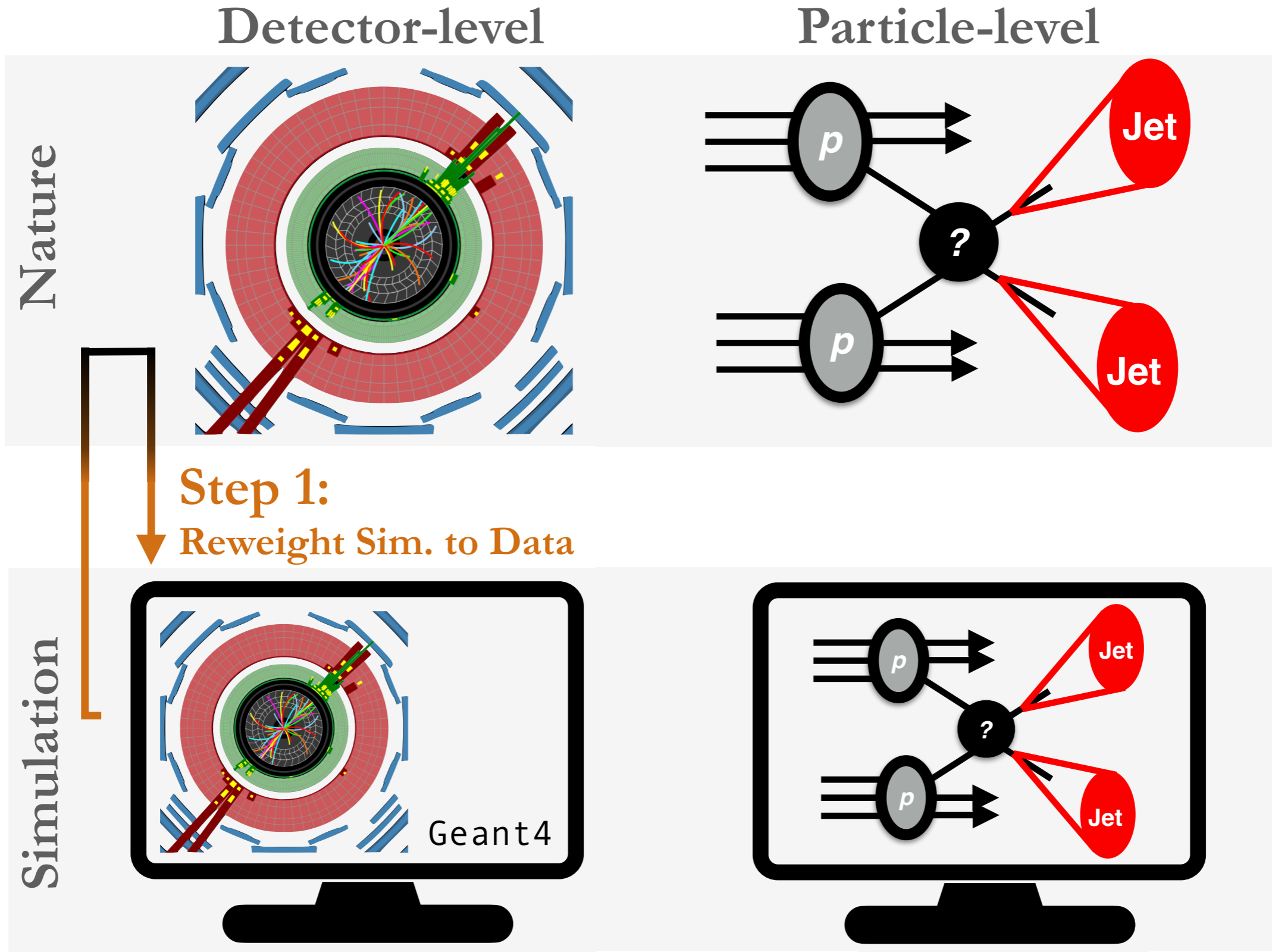
Particle-level



Simulation

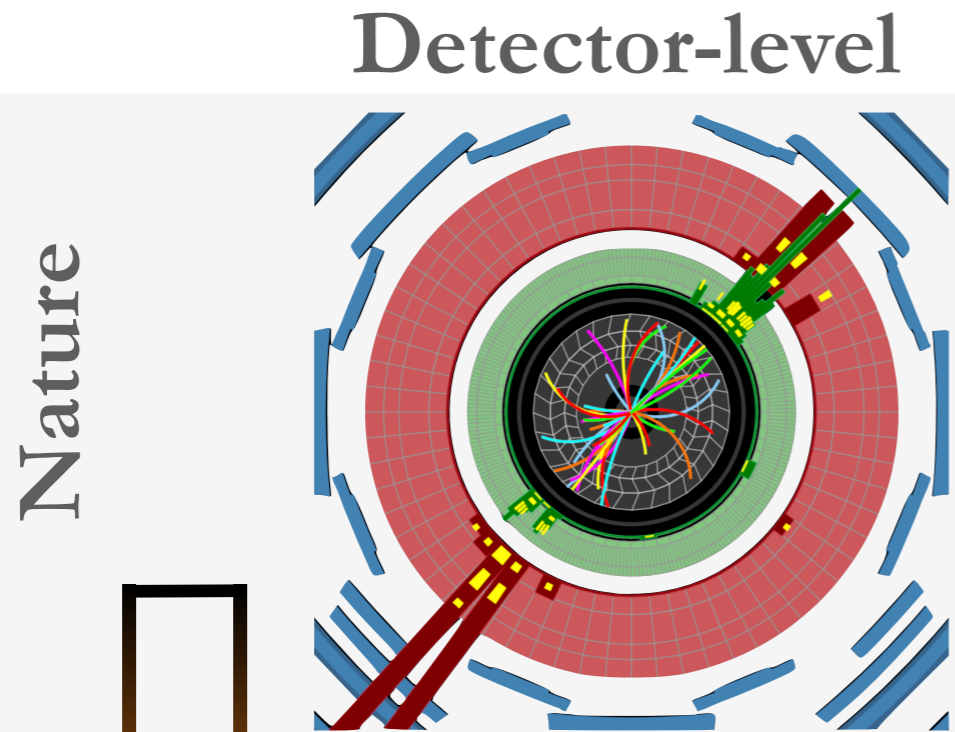


A brief introduction to OmniFold



A brief introduction to OmniFold

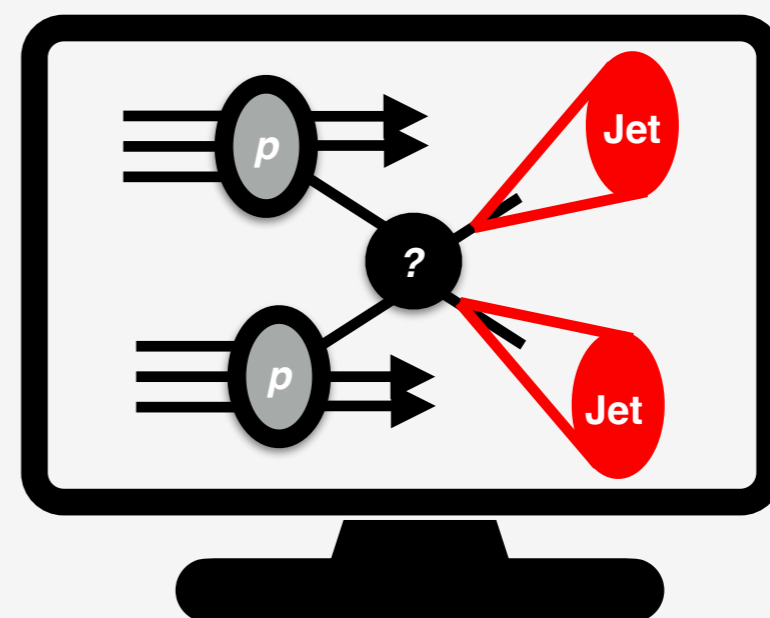
25



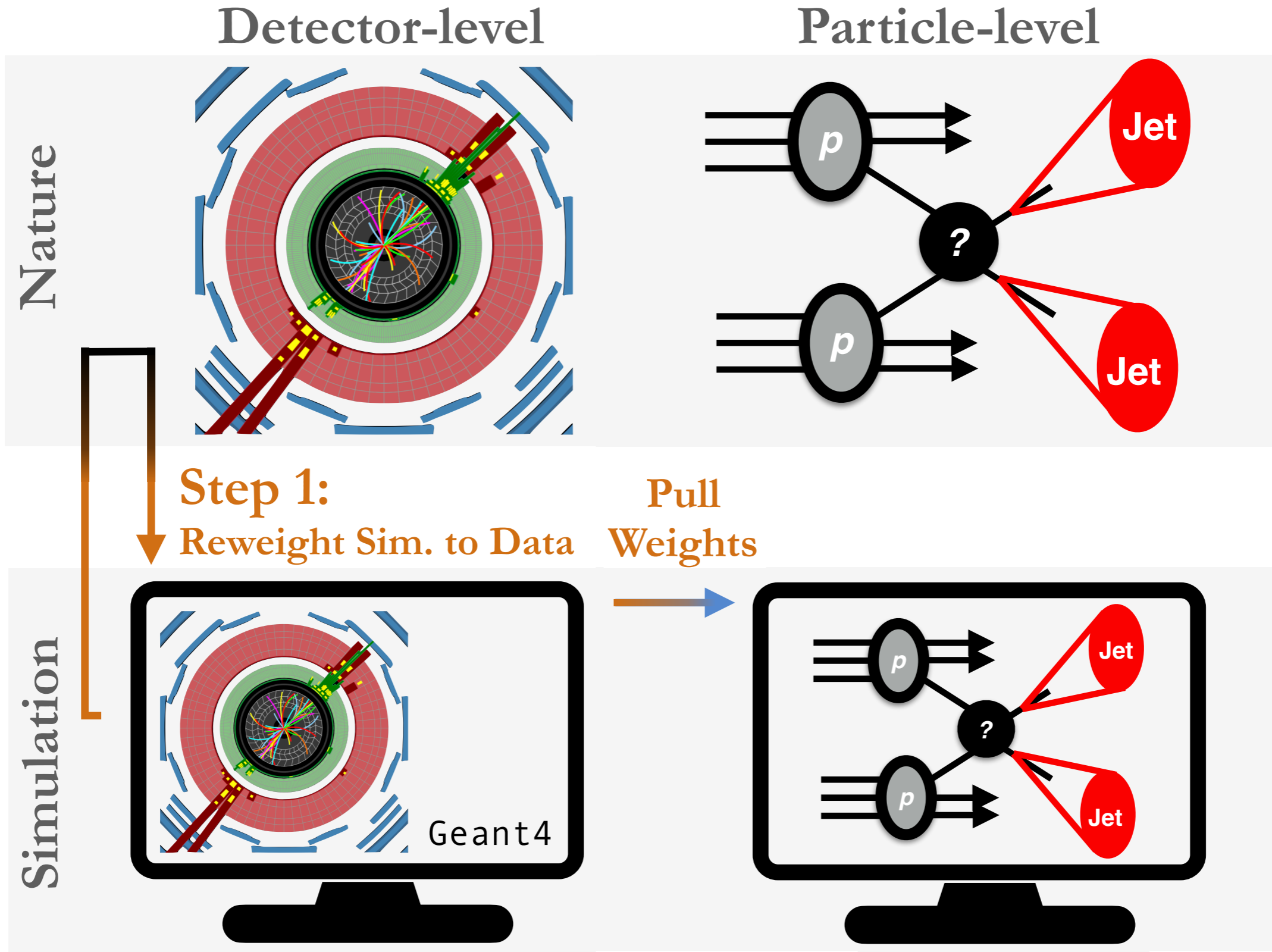
Particle-level

Unbinned, high-dimensional reweighting performed with neural networks

Step 1:
Reweight Sim. to Data

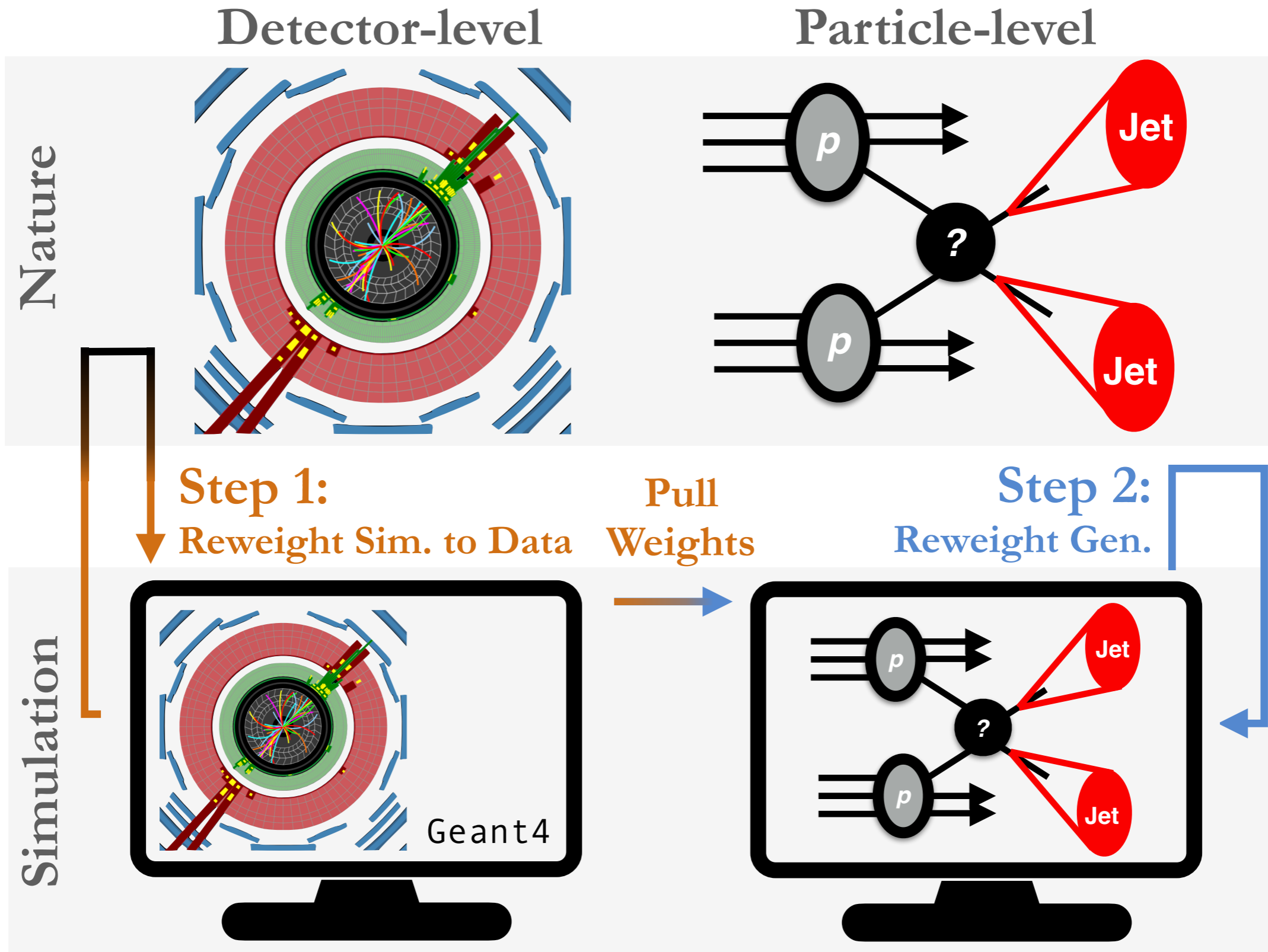


A brief introduction to OmniFold

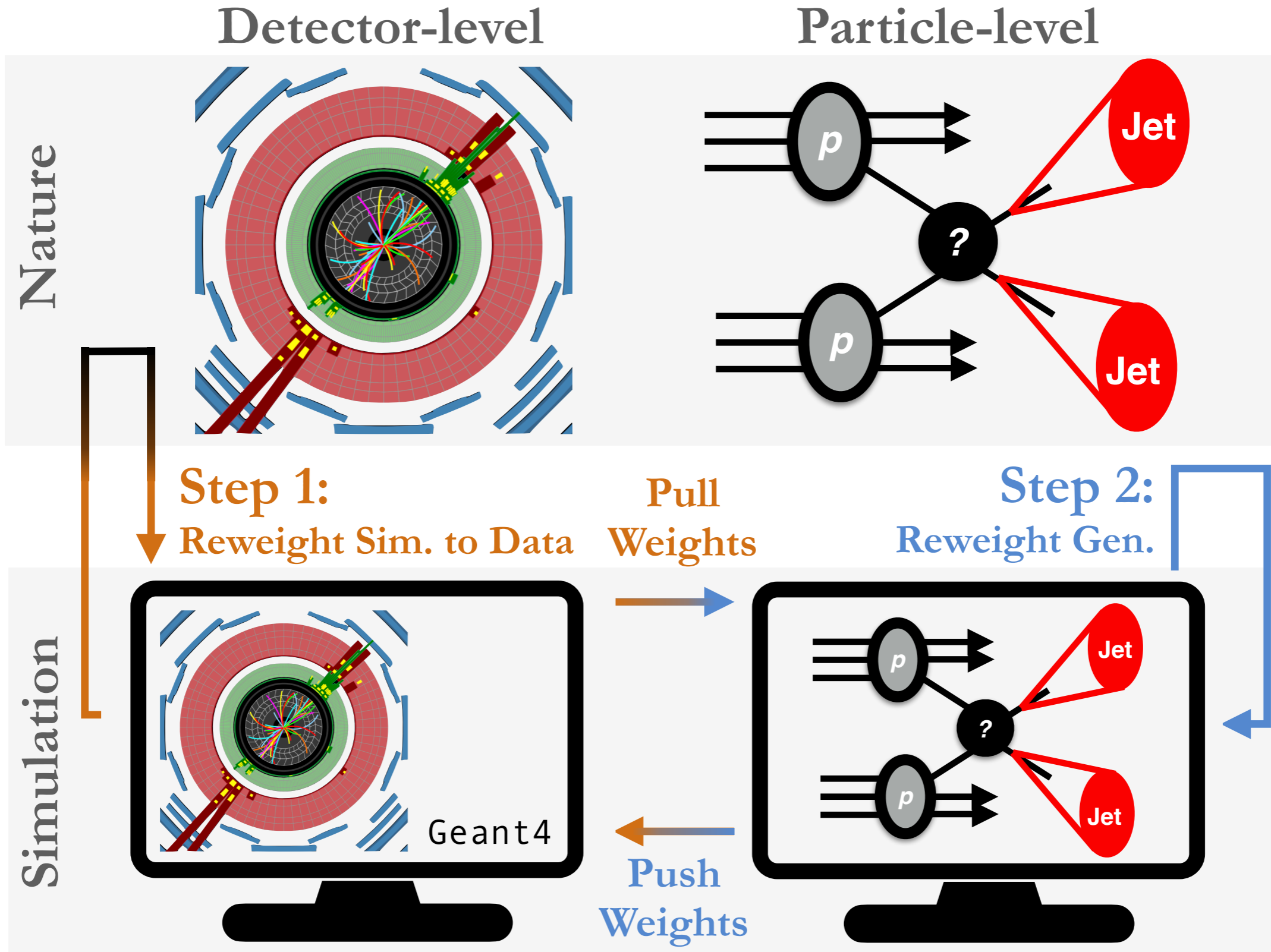


A brief introduction to OmniFold

27

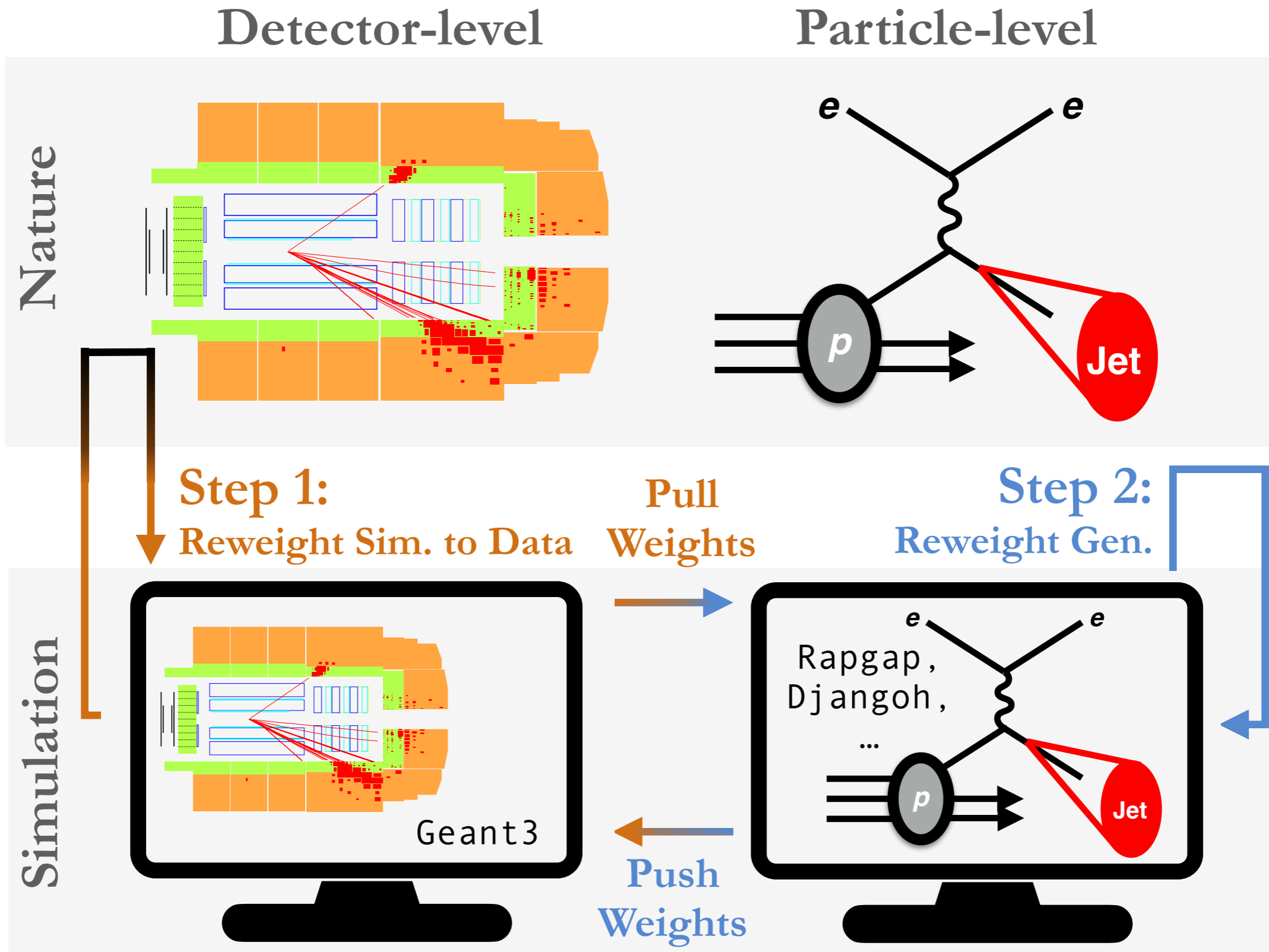


A brief introduction to OmniFold



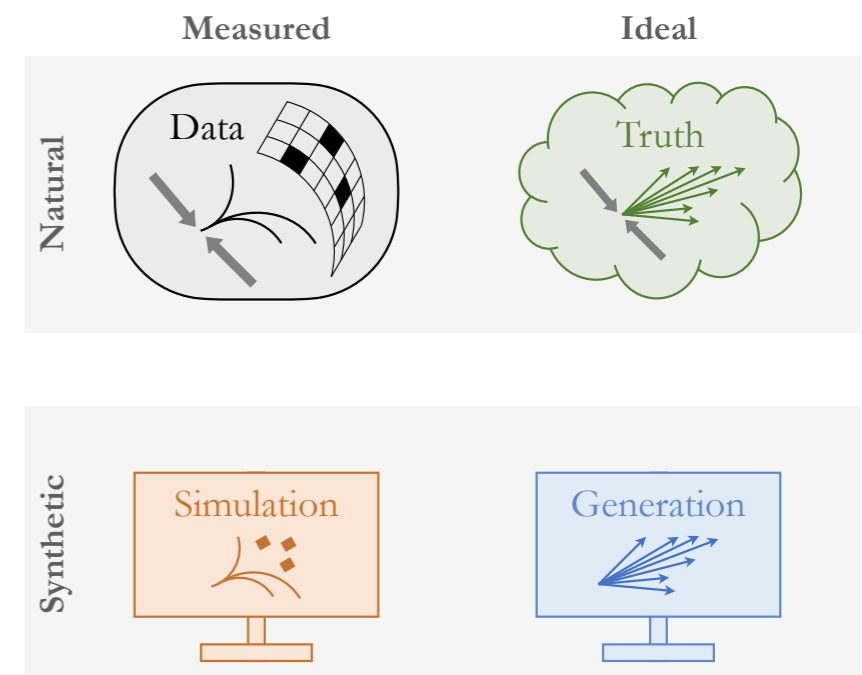
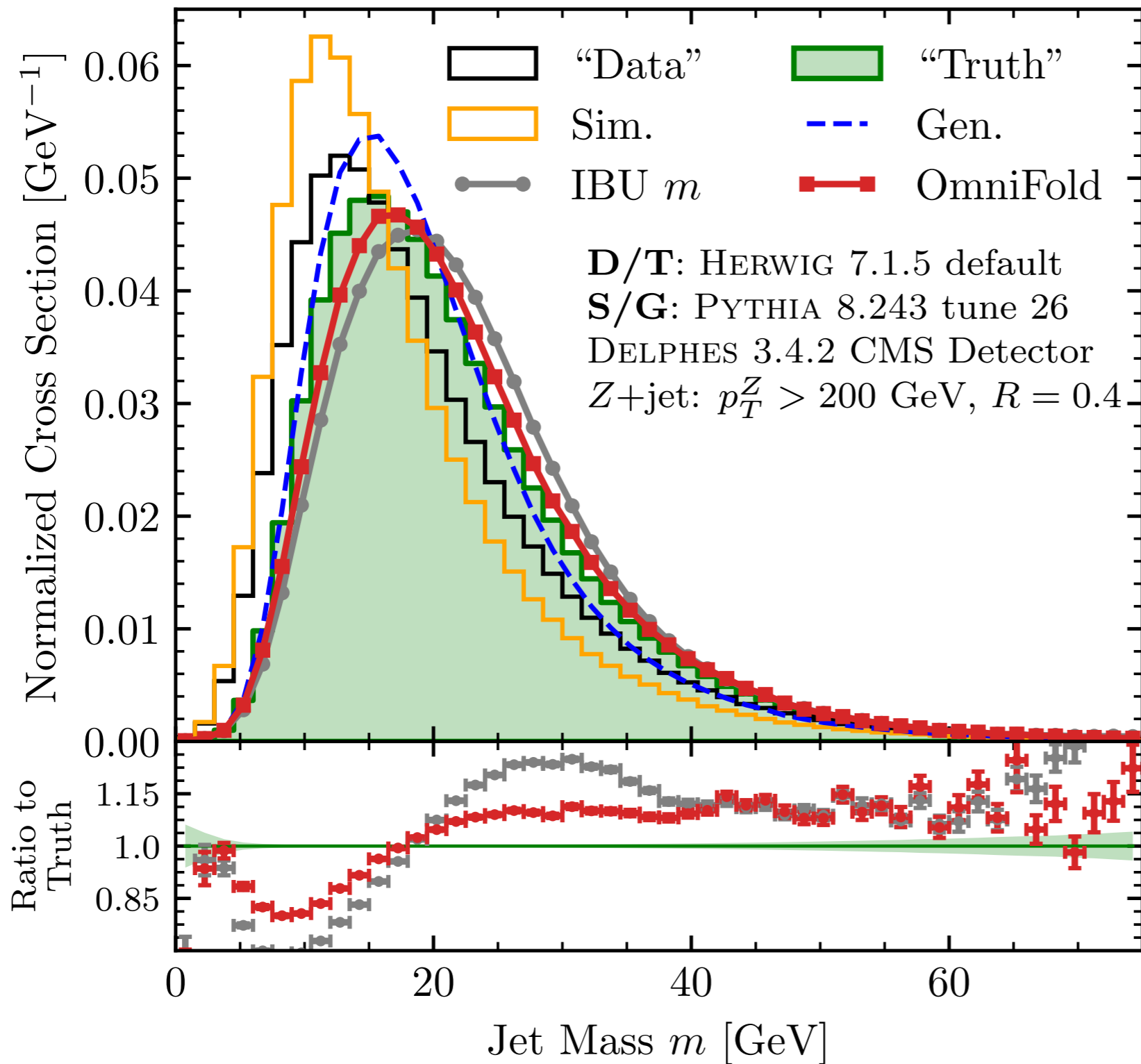
A brief introduction to OmniFold

29



Full phase-space unfolding

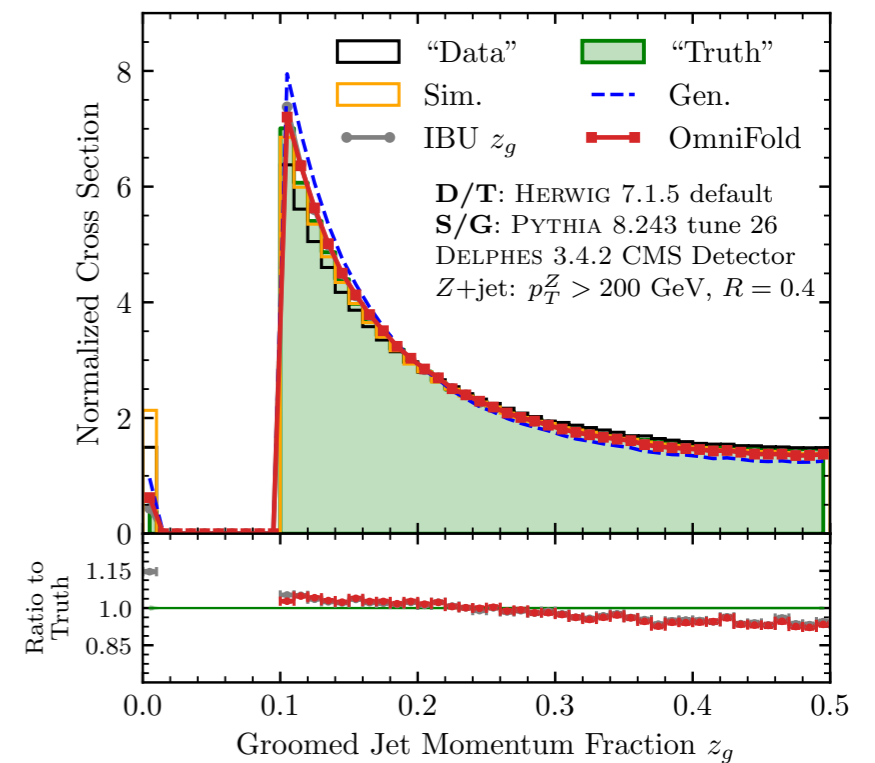
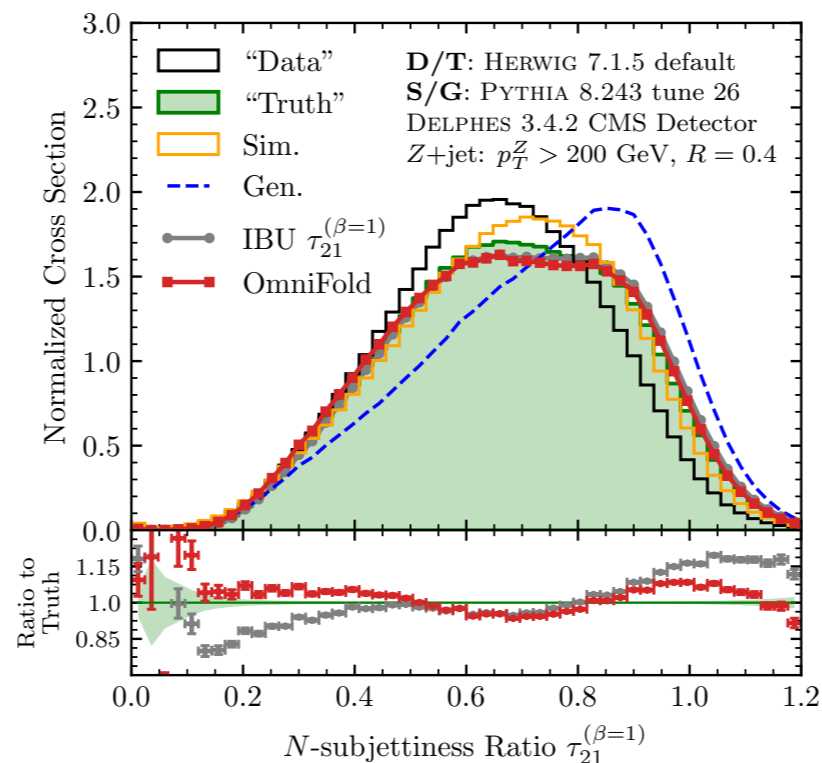
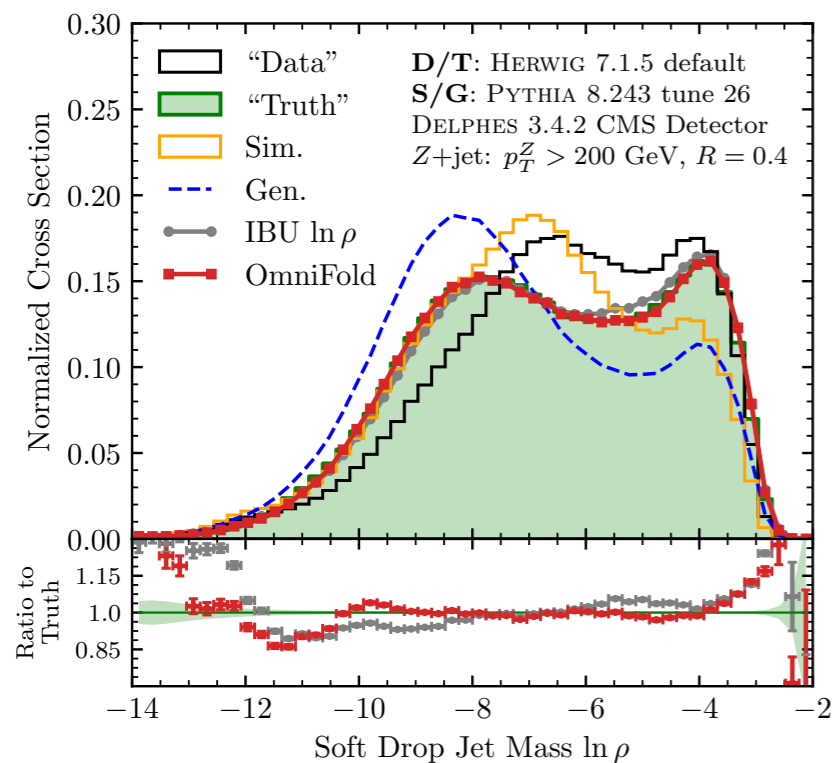
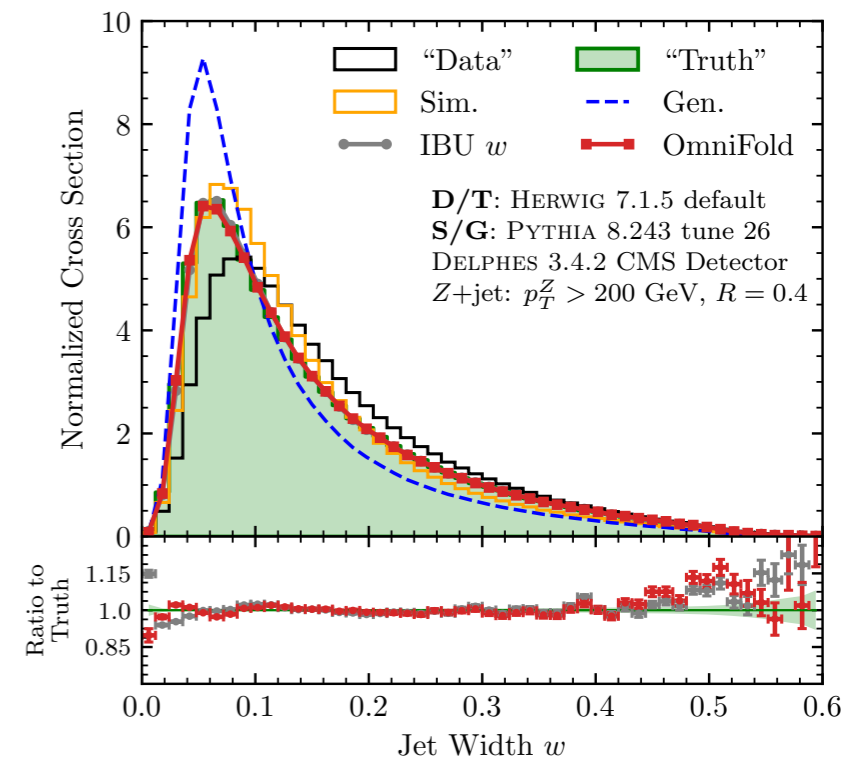
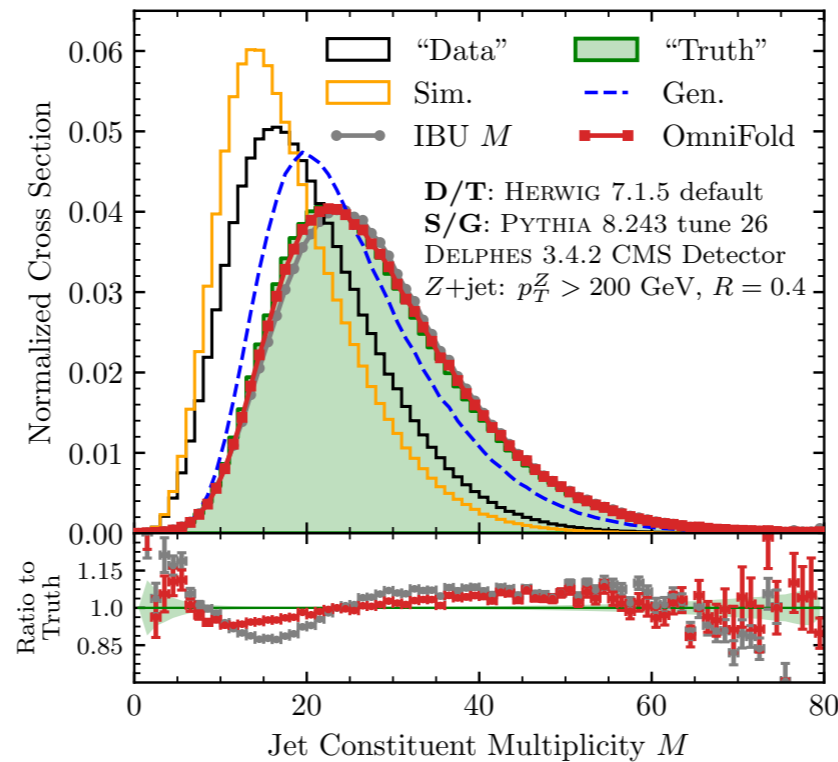
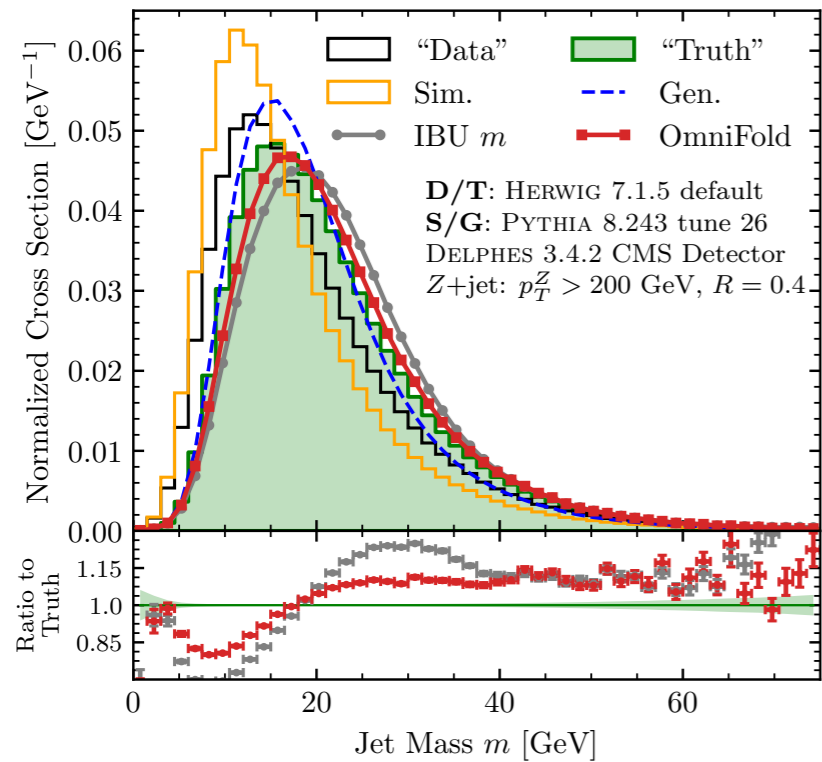
A. Andreassen, P. Komiske, E. Metodiev, BPN, J. Thaler, PRL 124 (2020) 182001



Full phase-space unfolding

31

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



Full phase-space unfolding

32

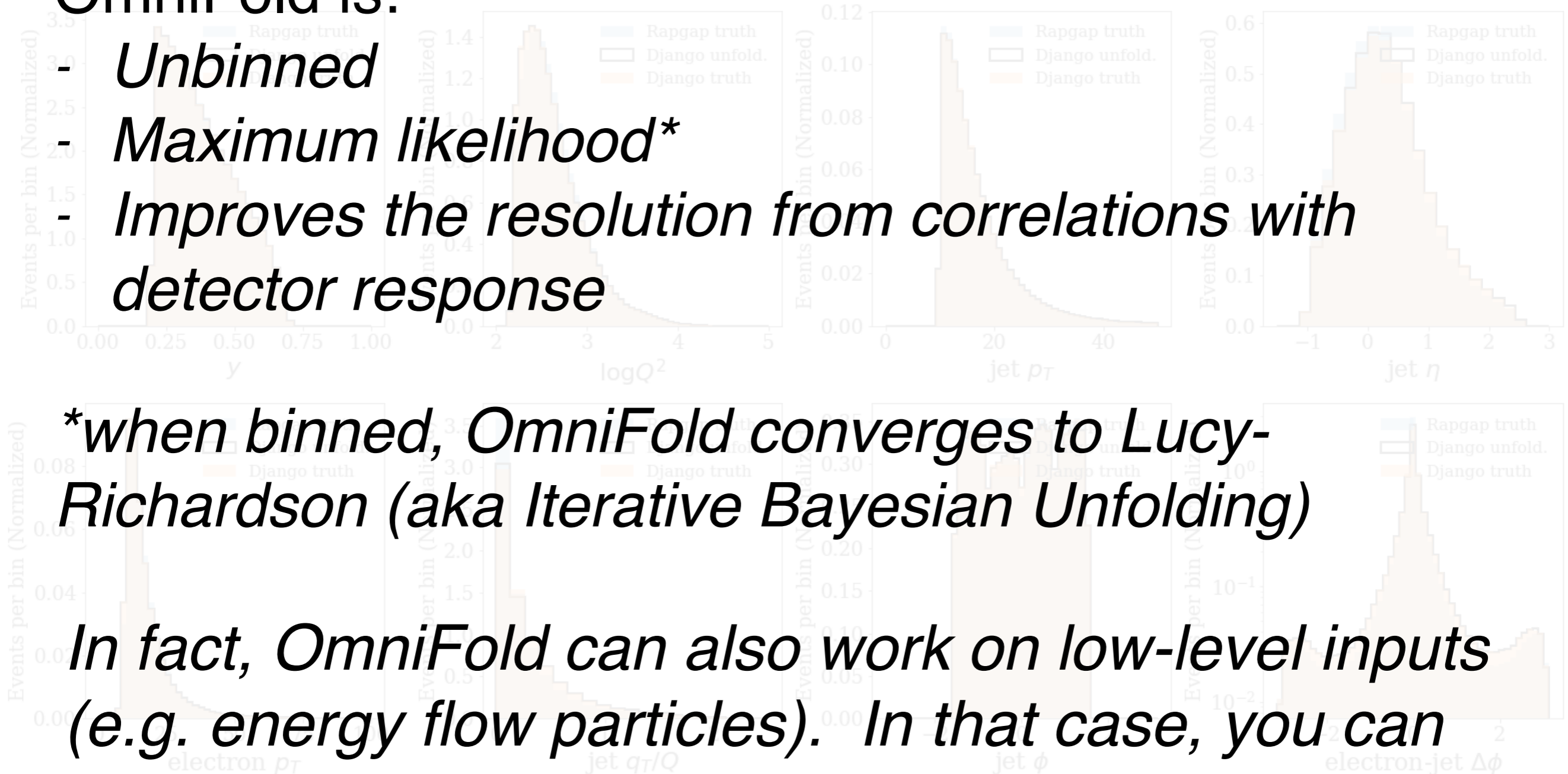
OmniFold is:

- *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.*

We see excellent closure for the full phase space!



→ Physics details in Miguel's talk

Some technical details

33



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>]



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, ...

35

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

Measurement of lepton-jet correlation in deep-inelastic scattering with the H1 detector using machine learning for unfolding

V. Andreev,²³ M. Arratia,³⁵ A. Bagdasaryan,⁴⁶ A. Baty,¹⁶ K. Begzsuren,³⁹ A. Belousov,²³, V. Boudry,³¹ G. Brandt,¹³ D. Britzger,²⁶ A. Buniatyán,⁶ L. Bystritskaya,²² A.J. Campbell,¹⁴ K.B. Chen,³¹ J.G. Contreras,⁴¹ J. Cvach²⁷, J.B. Dainton¹⁹, K. Daum³⁰, A. Deshpande^{33,36}, C. I. Eckerlin¹⁴, S. Egli³⁷, E. Elsen¹⁴, L. Favart⁴, A. Fedotov⁴², J. Feltesse¹², M. Fleischer¹⁴, A. Fomenko²³, C. Gal³⁸, J. Gayler¹⁴, L. Goerlich¹⁷, N. Gogitidze²³, M. Grab¹⁹, T. Greenshaw¹⁹, G. Grindhammer²⁶, D. Haidt¹⁴, R.C.W. Henderson¹⁸, J. Hessler²⁶, D. Hoffmann²¹, R. Horisberger⁴³, T. Hreus⁵⁰, F. Huber¹⁵, P.M. Jacobs⁵, M. Jacquet²⁹, T. Janssen¹⁴, H. Jung¹⁴, M. Kapichine¹⁰, J. Katzy¹⁴, C. Kiesling²⁶, M. Klein¹⁹, C. Kleinwort¹⁴, H.T. Klest³⁸, P. Kostka¹⁹, J. Kretschmar¹⁹, D. Krücker¹⁴, K. Krüger¹⁴, M.P.J. Landon²⁰, W. Lange⁴⁸, P. I. S.H. Lee³, S. Levonian¹⁴, W. Li¹⁶, J. Lin¹⁶, K. Lipka¹⁴, B. List¹⁴, J. List¹⁴, B. Lobodzinski²⁶, E. H.-U. Martyn¹, S.J. Maxfield¹⁹, A. Mehta¹⁹, A.B. Meyer¹⁴, J. Meyer¹⁴, S. Mikocki¹⁷, M.M. Mondal³³, K. Müller⁵⁰, B. Nachman⁵, Th. Naumann⁴⁸, P.R. Newman⁶, C. Niebuhr¹⁴, G. Nowak¹⁷, J.E. D. Ozerov⁴³, S. Park³⁸, C. Pascaud²⁹, G.D. Patel¹⁹, E. Perez¹¹, A. Petrukhin⁴², I. Picuric³², R. Polifka³⁴, S. Preins³⁵, V. Radescu³⁰, N. Raicevic³², T. Ravdandorj³⁹, P. Reimer³³, E. Rizvi²⁰, P. R. Roosen⁴, A. Rostovtsev²⁵, M. Rotaru⁷, D.P.C. Sankey⁹, M. Sauter¹⁵, E. Sauvan^{21,2}, S. Sch. B.A. Schmookler³⁸, L. Schoeffel¹², A. Schöning¹⁵, F. Sefkow¹⁴, S. Shushkevich²⁴, Y. Soloviev²³, D. South¹⁴, V. Spaskov¹⁰, A. Specka³¹, M. Steder¹⁴, B. Stella³⁶, U. Straumann⁵⁰, C. Sun³⁷, T. P.D. Thompson⁶, D. Traynor²⁰, B. Tseepeldorj³⁹, Z. Tu⁴¹, A. Valkárová³⁴, C. Vallée²¹, P. Van D. Wegener⁹, E. Wünsch¹⁴, J. Žáček³⁴, J. Zhang³⁷, Z. Zhang²⁹, R. Zlebčik³⁴, H. Zohrabyan⁴⁶ and (The H1 Collaboration)

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

© 2022 CERN for the benefit of the LHCb collaboration. CC BY 4.0 licence.

Unbinned Deep Learning Jet Substructure Measurement in High Q^2 ep collisions at HERA

Andreev⁴⁴, M. Arratia²⁹, A. Bagdasaryan⁴⁰, A. Baty¹⁶, K. Begzsuren³⁴, A. Bolz¹⁴, V. Boudry²⁵, G. Brandt¹³, Britzger²², A. Buniatyán¹, L. Bystritskaya⁴⁴, A.J. Campbell¹⁴, K.B. Cantun Avila⁴¹, K. C. Chen³¹, J.G. Contreras⁴¹, J. Cvach²⁷, J.B. Dainton¹⁹, K. Daum³⁰, A. Deshpande^{33,36}, C. I. Eckerlin¹⁴, S. Egli³⁷, E. Elsen¹⁴, L. Favart⁴, A. Fedotov⁴², J. Feltesse¹², M. Fleischer¹⁴, J. Gayler¹⁴, L. Goerlich¹⁷, N. Gogitidze¹⁴, M. Gouzevitch²³, C. Grab²⁷, T. Greenshaw¹⁹, Haidt¹⁴, R.C.W. Henderson¹⁸, J. Hessler²², J. Hladky²⁷, D. Hoffmann³¹, R. Horisberger²⁷, P.M. Jacobs⁵, M. Jacquet²⁴, T. Janssen⁴, A.W. Jung²⁸, J. Katzy¹⁴, C. Kiesling²², M. Klein¹⁹, T. Klest³³, R. Kogler¹⁴, P. Kostka¹⁹, J. Kretschmar¹⁹, D. Krücker¹⁴, K. Krüger¹⁴, M.P.J. Laycock²⁶, S.H. Lee³, S. Levonian¹⁴, W. Li¹⁶, J. Lin¹⁶, K. Lipka¹⁴, B. List¹⁴, J. List¹⁴, Long²⁹, E. Malinovskii⁴⁴, H.-U. Martyn¹, S.J. Maxfield¹⁹, A. Mehta¹⁹, A.B. Meyer¹⁴, J. V.M. Mikuni², M.M. Mondal³³, K. Müller⁵⁰, B. Nachman⁵, Th. Naumann¹⁴, P.R. Newn G. Nowak¹⁷, J.E. Olsson¹⁴, D. Ozerov⁴⁴, S. Park³³, C. Pascaud²⁹, G.D. Patel¹⁹, E. Perez I. Picuric²⁶, D. Pitzl¹⁴, R. Polifka²⁸, S. Preins²⁹, V. Radescu¹⁵, N. Raicevic²⁶, T. Ravdandorj²⁰, P. Robmann⁴³, R. Roosen⁴, A. Rostovtsev⁴⁴, M. Rotaru⁷, D.P.C. Sankey⁹, M. S. S. Schmitt¹⁴, B.A. Schmookler³³, G. Schnell¹², L. Schoeffel¹², A. Schöning¹⁵, F. Sefkow²³, oloviev¹⁴, P. Sopic¹⁷, D. South¹⁴, A. Specka²⁵, M. Steder¹⁴, B. Stella³⁰, U. Straumann⁶, P.D. Thompson⁷, F. Torales Acosta⁴, D. Traynor²⁰, B. Tseepeldorj^{34,35}, Z. Tu³⁶, G. Tusti C. Vallée²¹, P. Van Mechelen¹⁰, D. Wegener¹⁰, E. Wünsch¹⁴, J. Žáček²⁸, J. Zhang³¹, Z. Zhang H. Zohrabyan⁴⁰, F. Zomer²⁴

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, Multi-Fold, 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:2108.12376v2 [hep-ex] 1 Apr 2022

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

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

+CMS open data study

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:

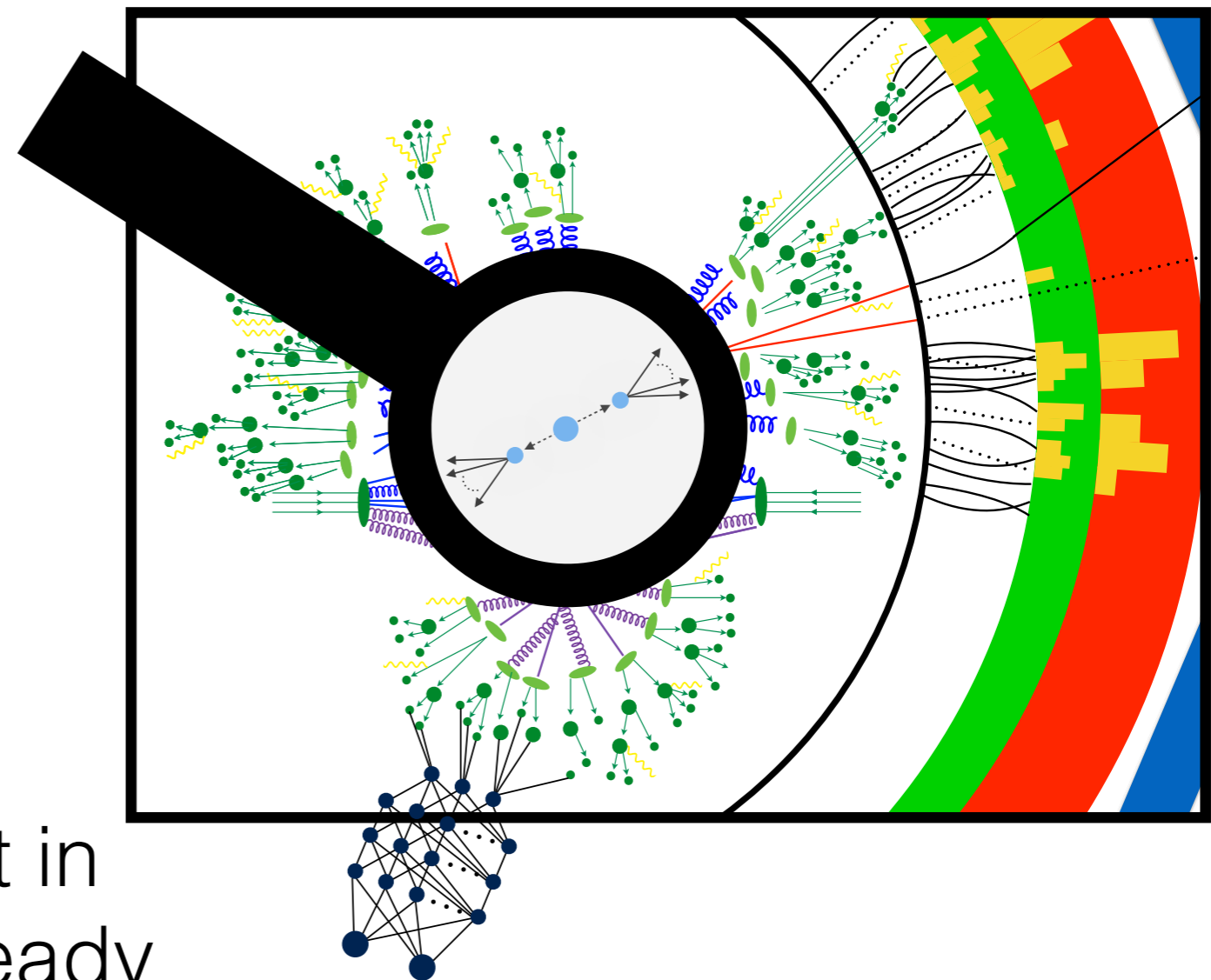
- OmniFold is computationally expensive (need to train many networks, especially with ensembling to reach precision)
- How to publish an unbinned result? (all results so far are presented as binned) - see 2109.13243. Breaks HEPData!
- Modeling/closure uncertainties in high dimensions (not a new problem,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

Conclusions and Outlook

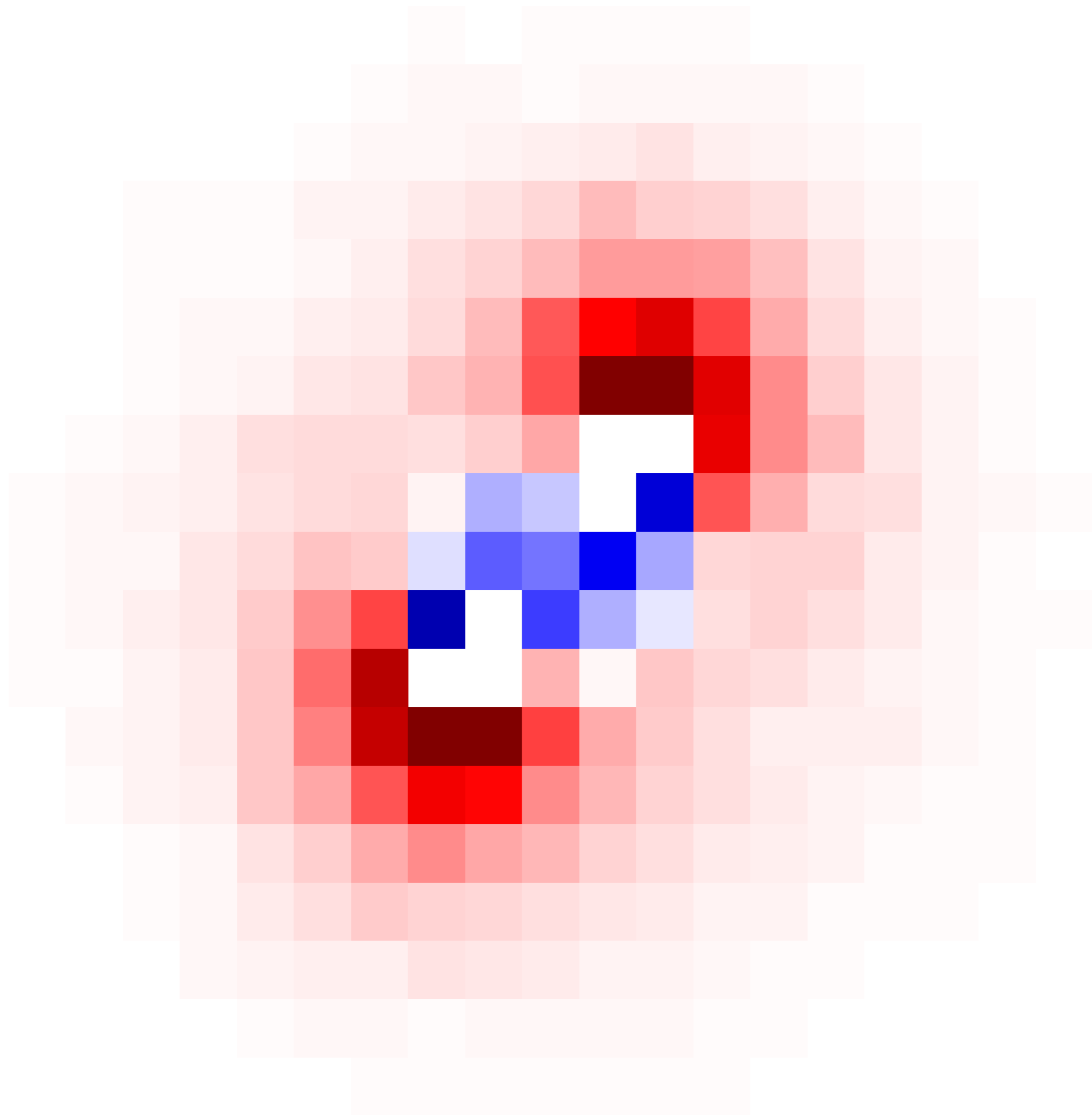
37

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



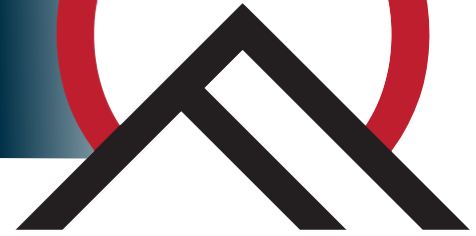
More R&D is required, but in parallel, these tools are already starting to **deliver science results!**



Fin.

Reweighting

39



How do to the reweighting without binning?

How do to the reweighting without binning?

dataset 1: sampled from $p(x)$

dataset 2: sampled from $q(x)$

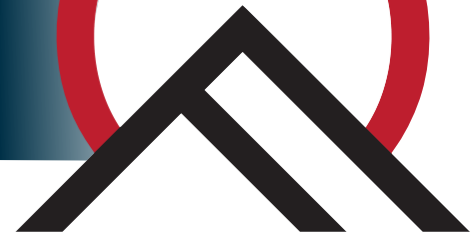
Create weights $w(x) = q(x)/p(x)$ so that when dataset 1 is weighted by w , it is statistically identical to dataset 2.

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

41



Fact: Neural 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!

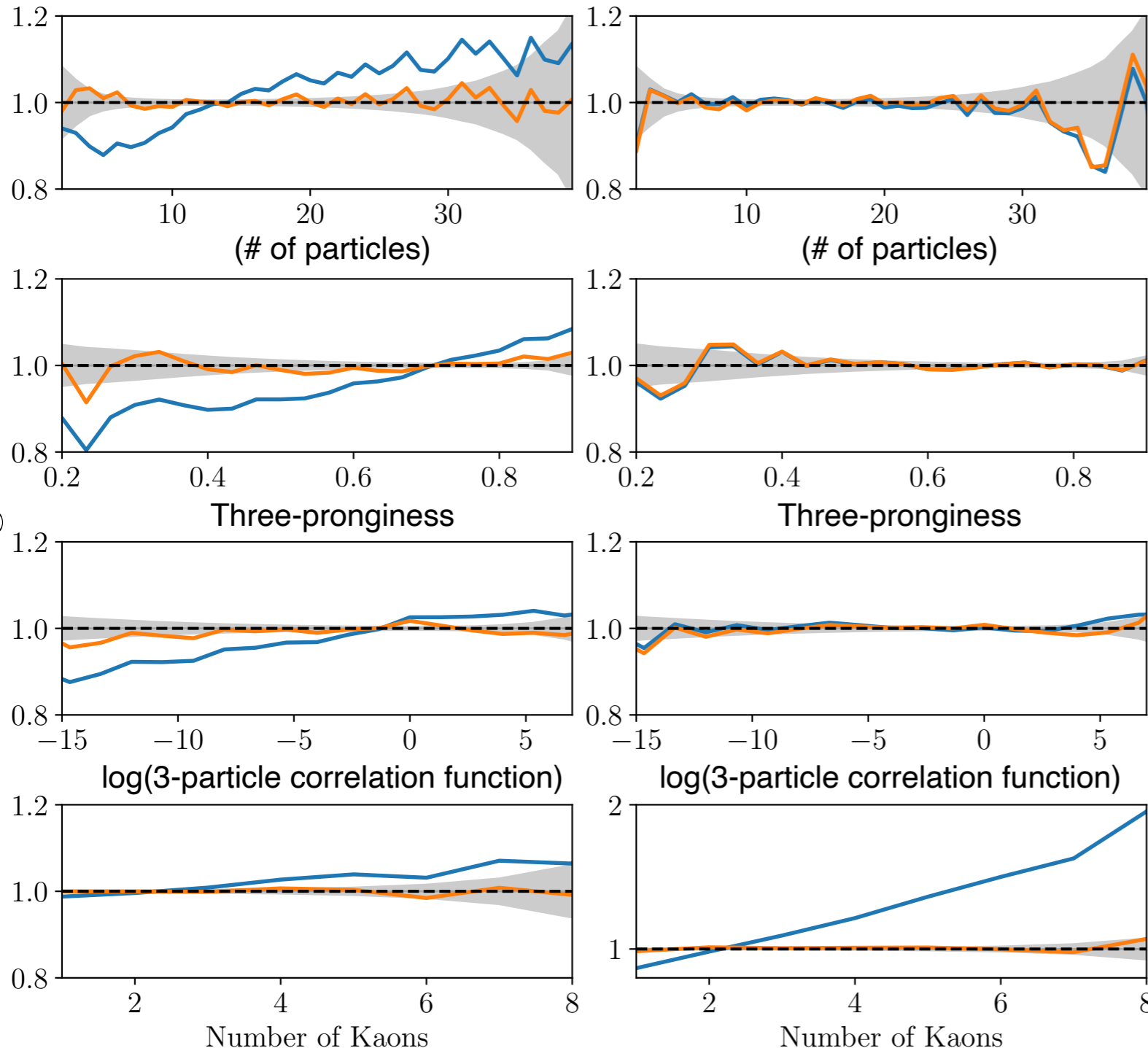
42



StringZ:aLund

StringFlav:probStoUD

— Unweighted — Weighted

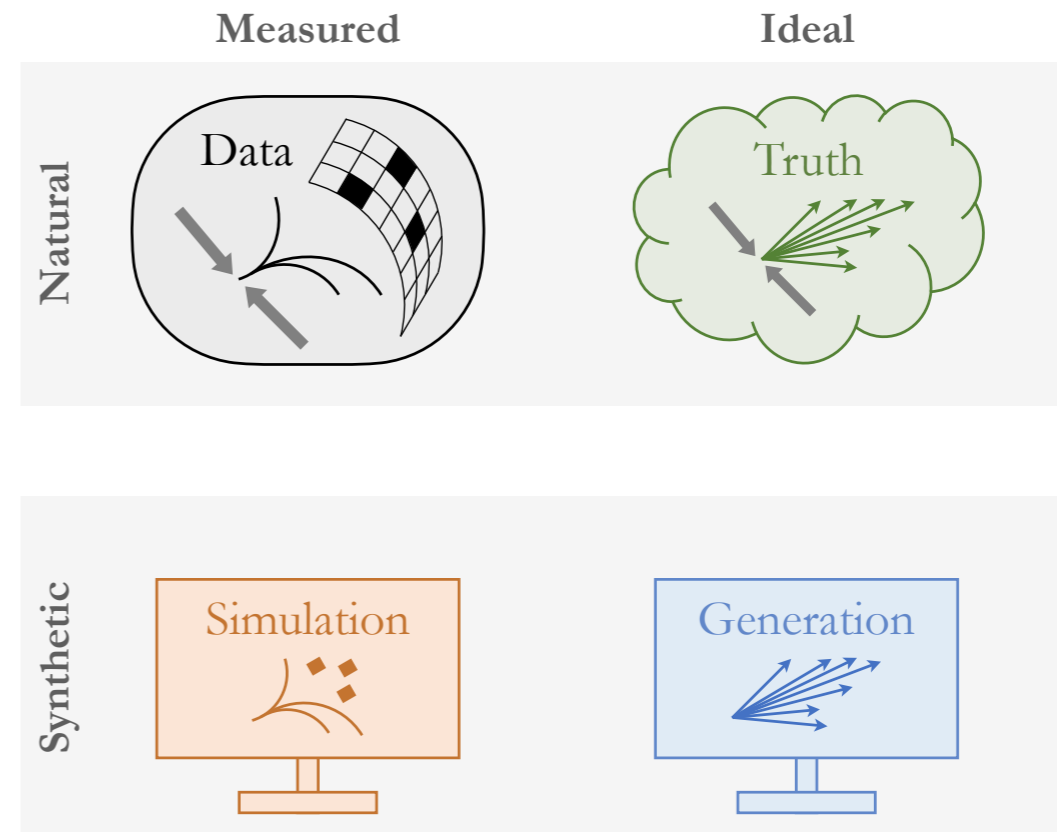
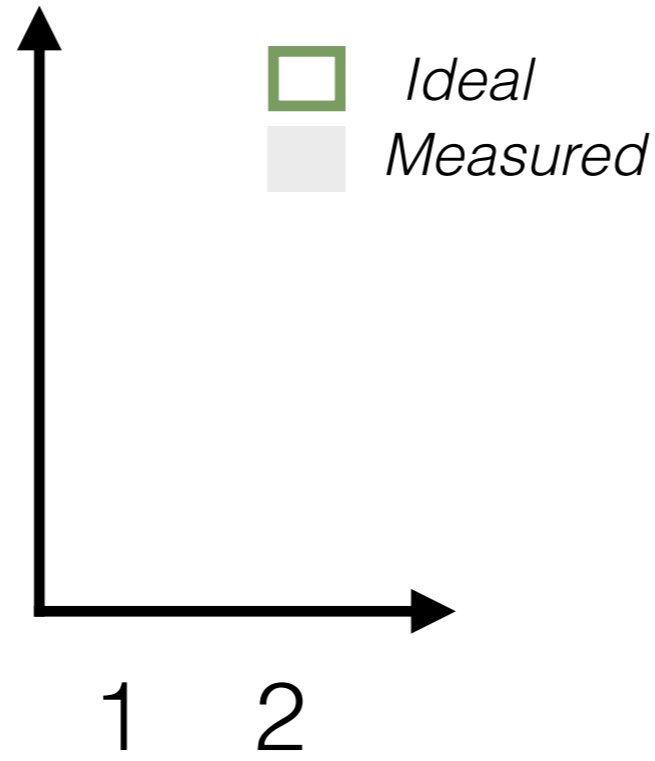
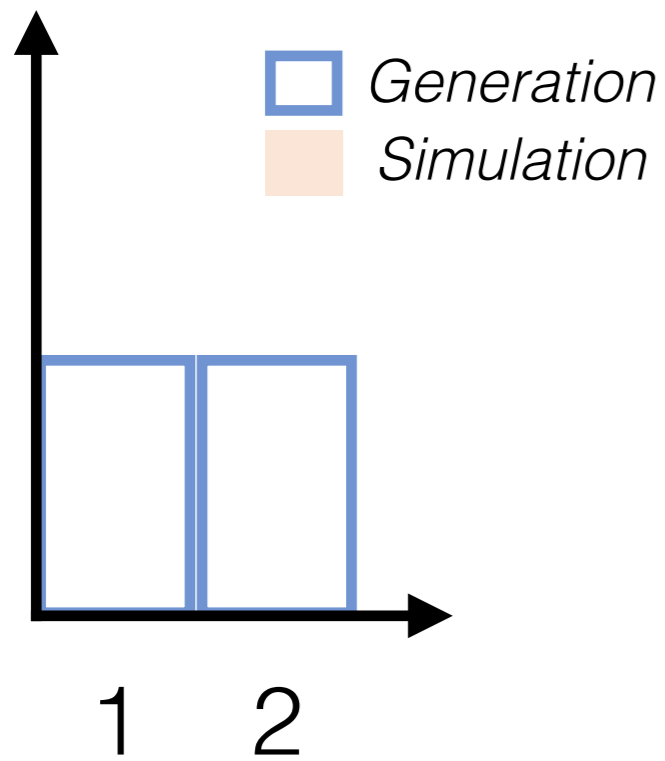


Full phase-space reweighing using simulated e^+e^-

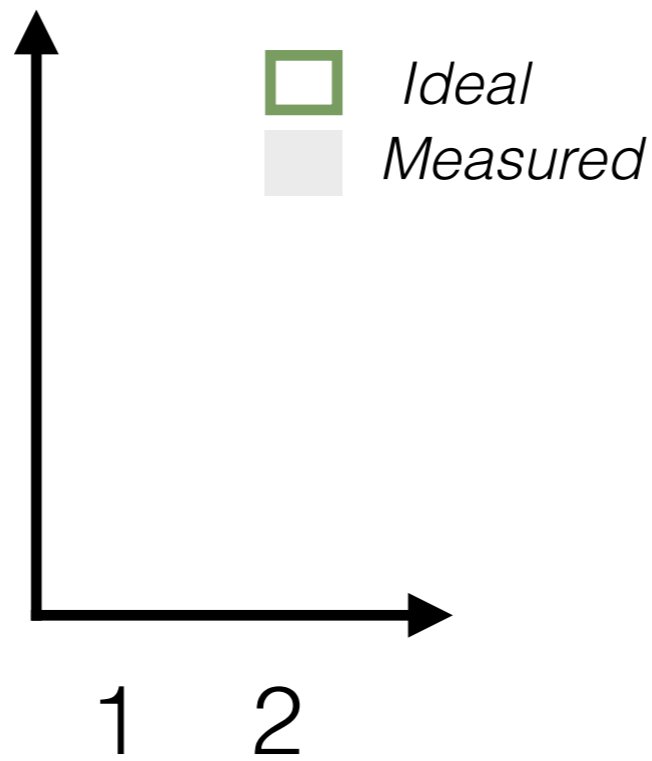
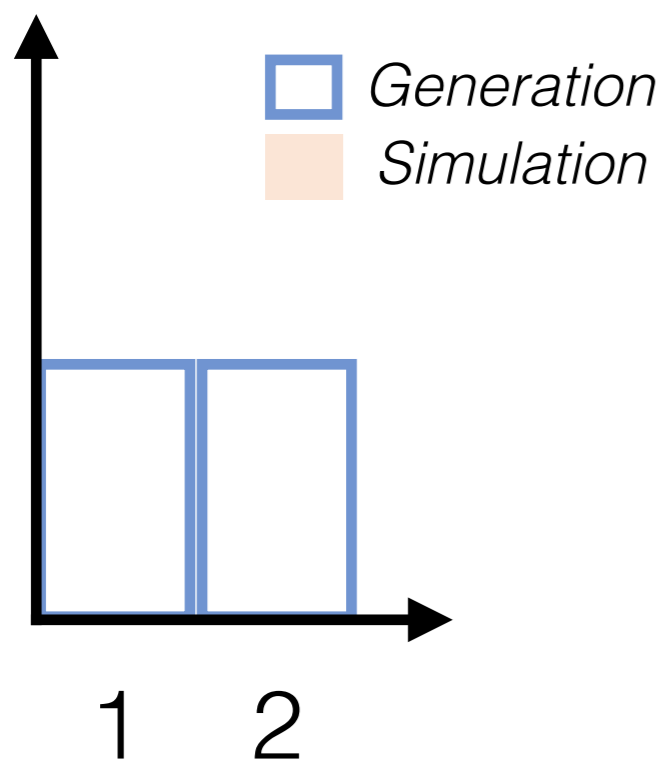
Works even when the differences are **small** (left) or **localized** (right).

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

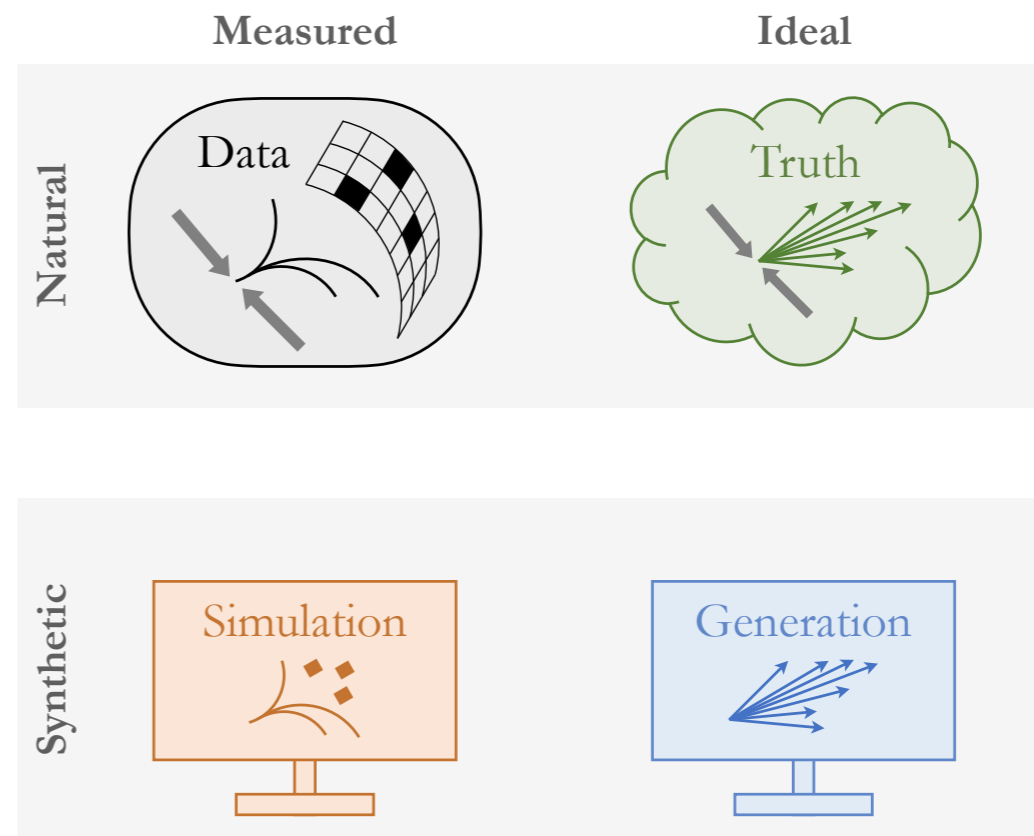
Unfold by iterating: OmniFold



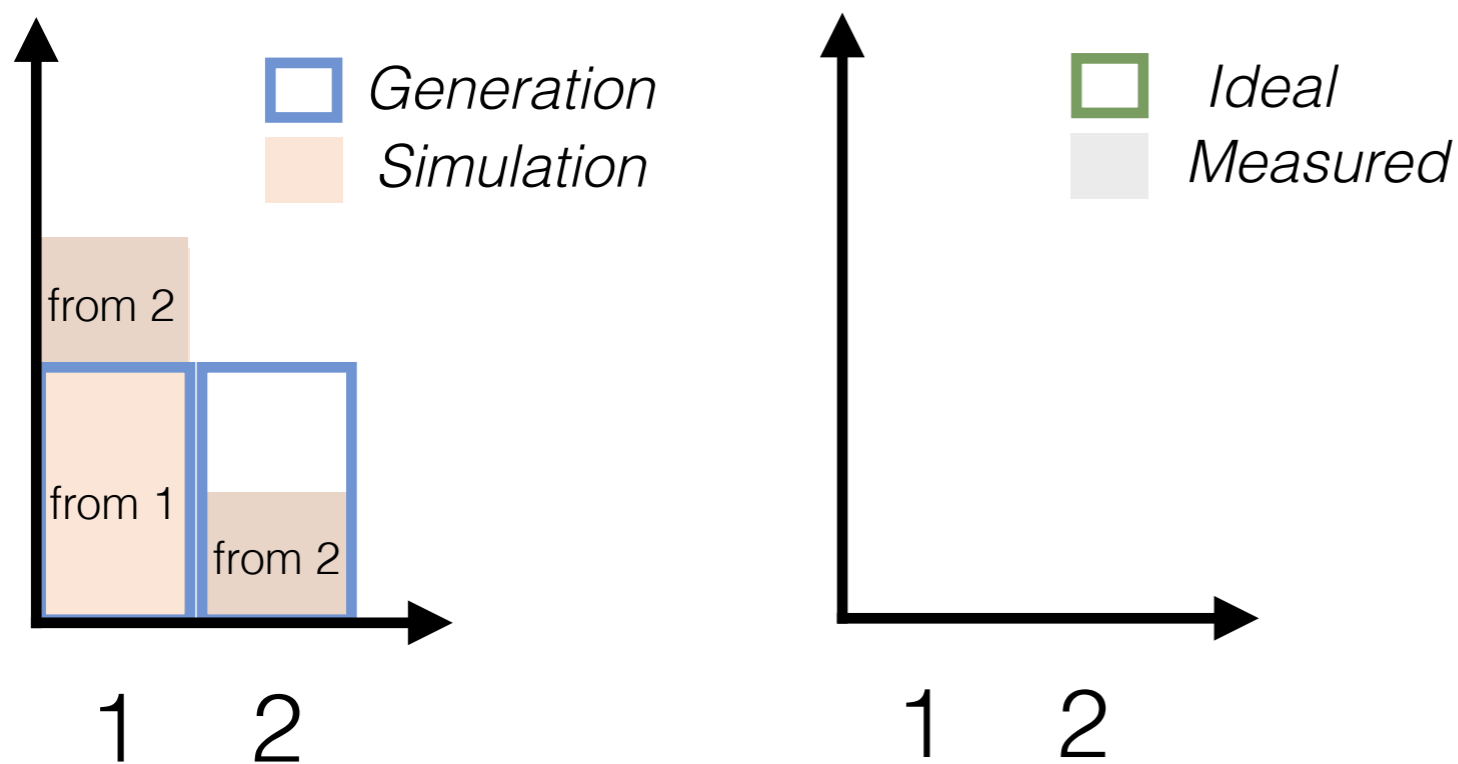
Unfold by iterating: OmniFold



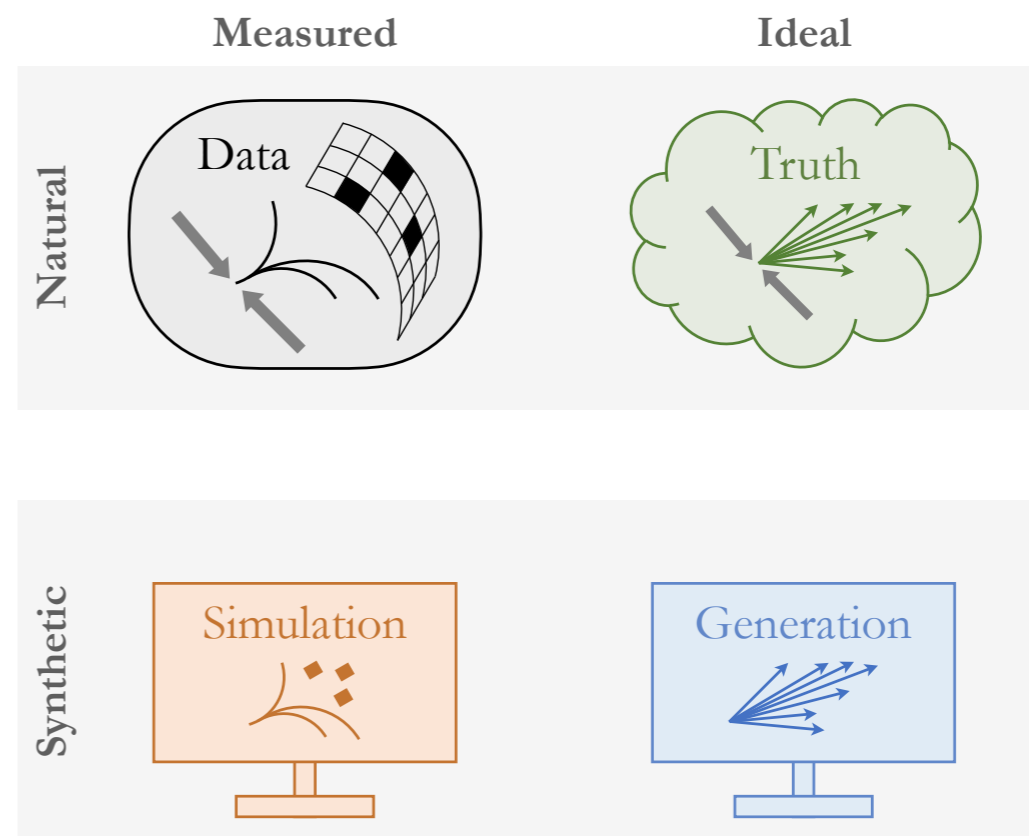
Measured	2	0%	50%
	1	100%	50%
		1	2
		Ideal	



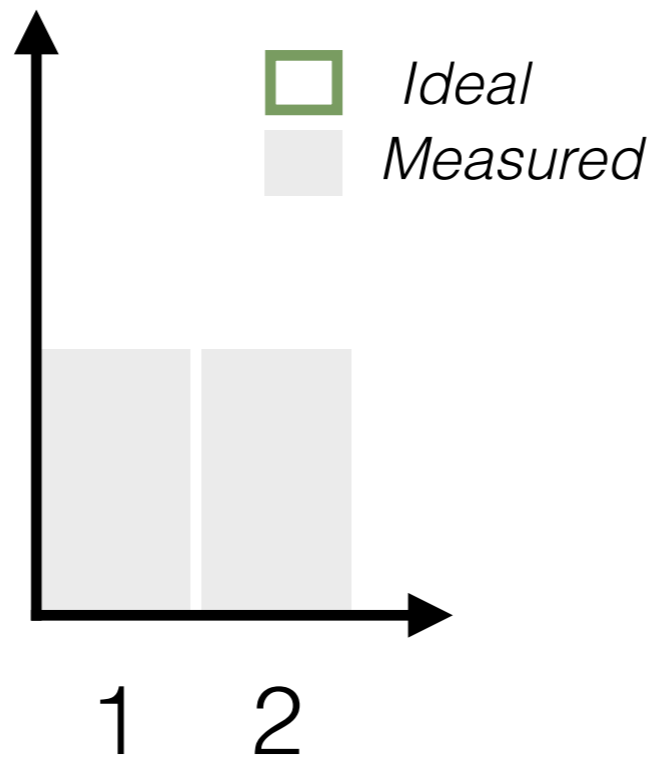
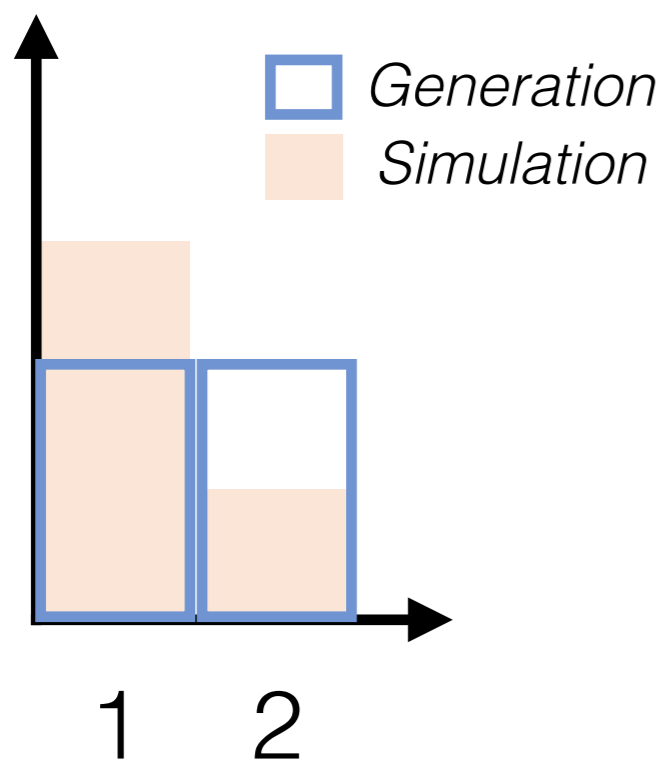
Unfold by iterating: OmniFold



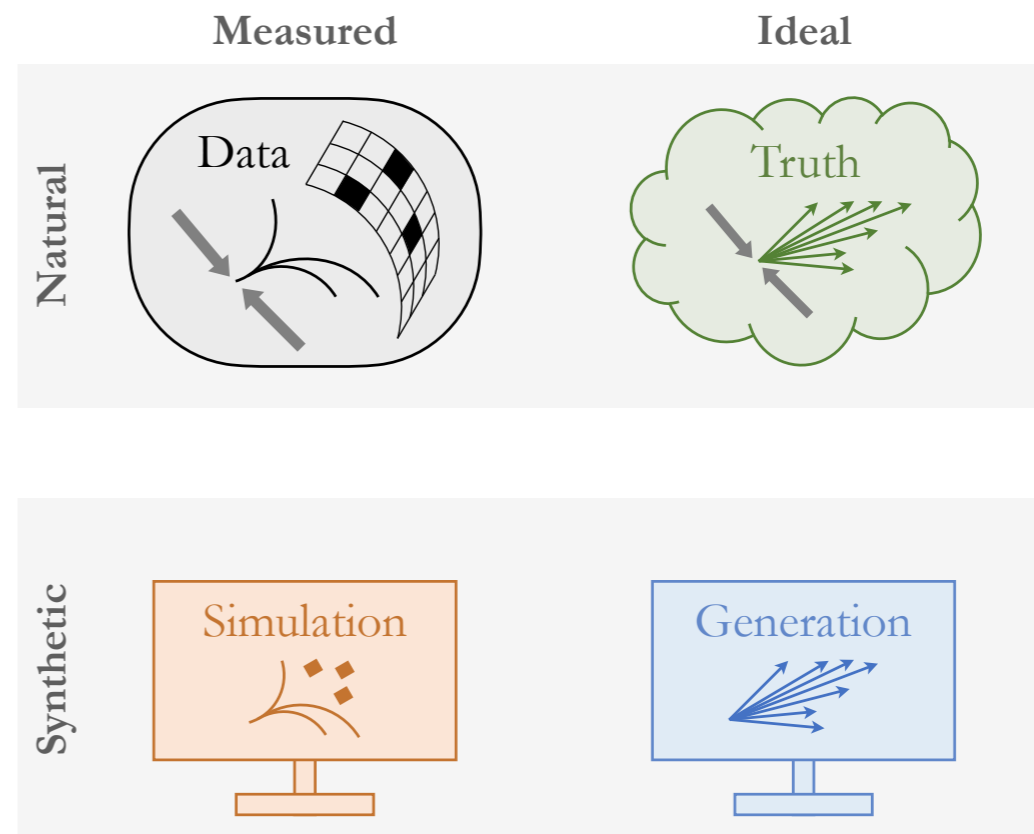
Measured	2	0%	50%
	1	100%	50%
		1	2
		Ideal	



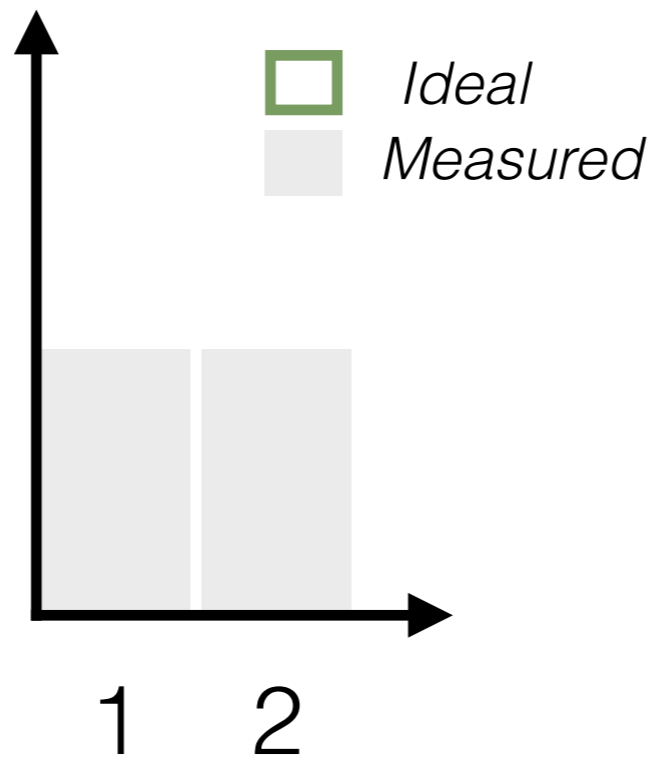
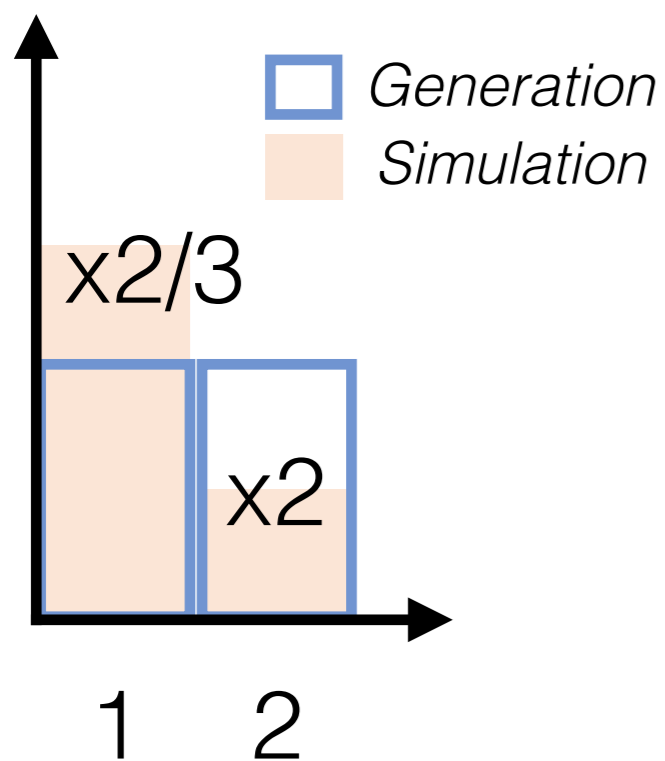
Unfold by iterating: OmniFold



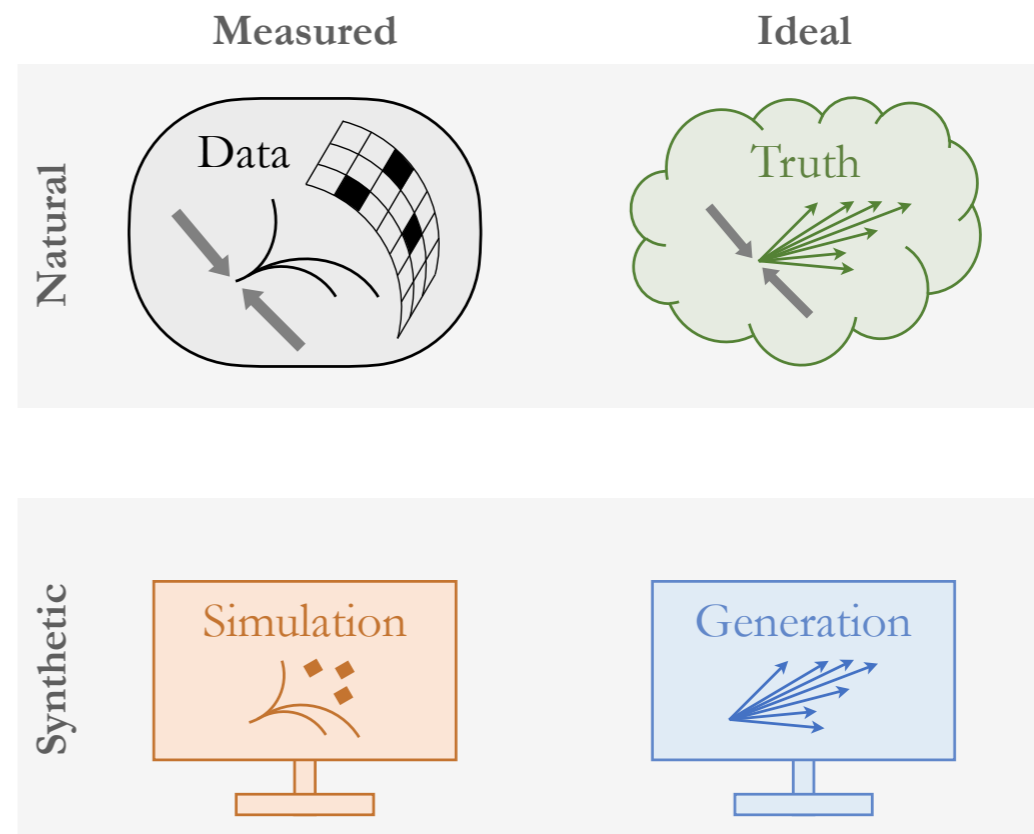
Measured	2	0%	50%
	1	100%	50%
		1	2
		Ideal	



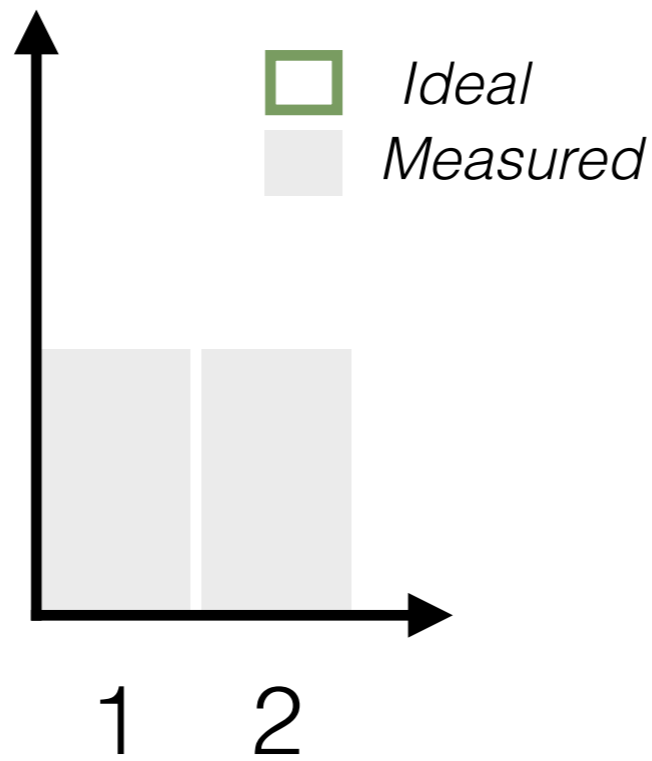
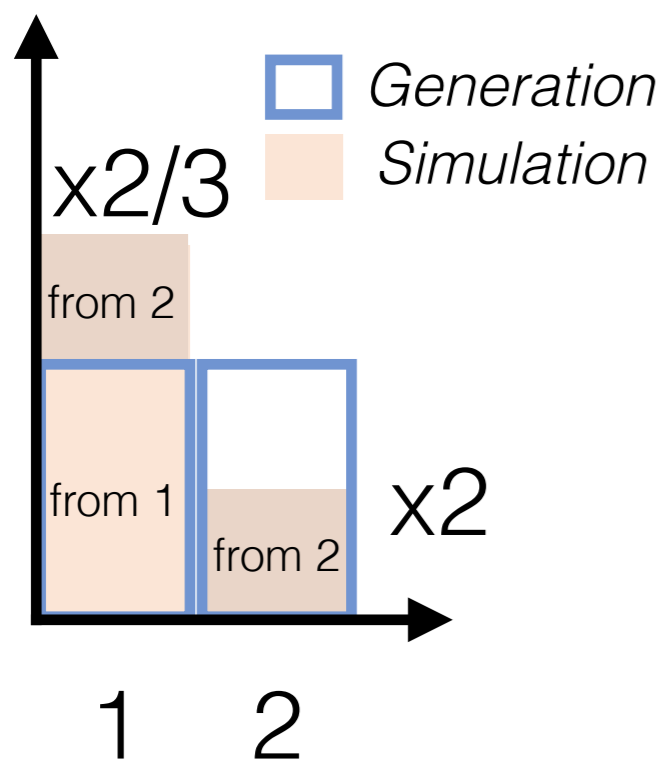
Unfold by iterating: OmniFold



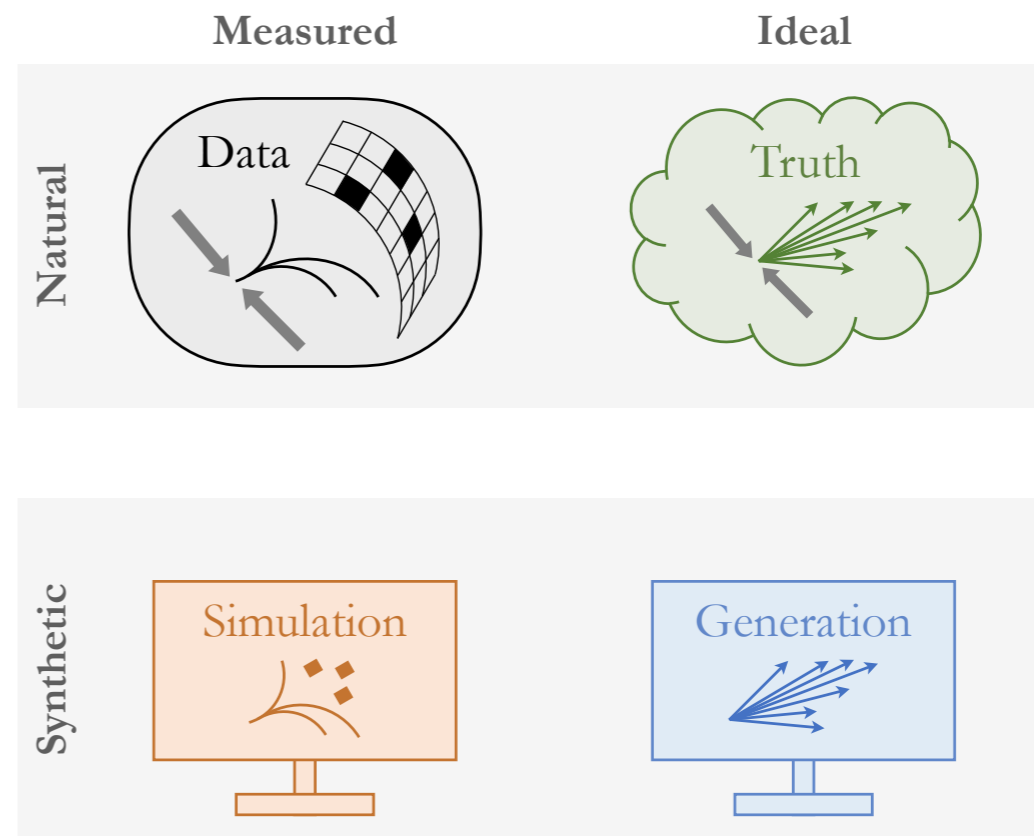
Measured	2	0%	50%
	1	100%	50%
		1	2
		Ideal	



Unfold by iterating: OmniFold

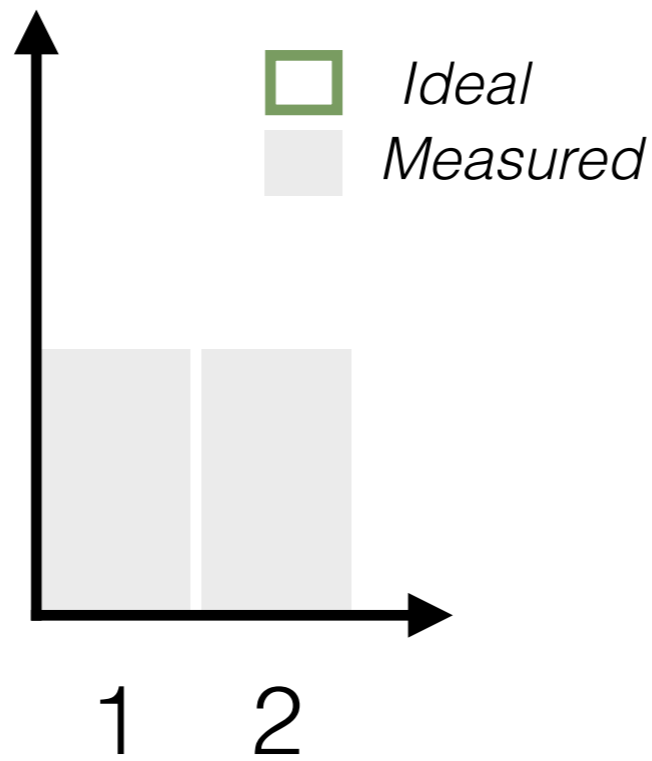
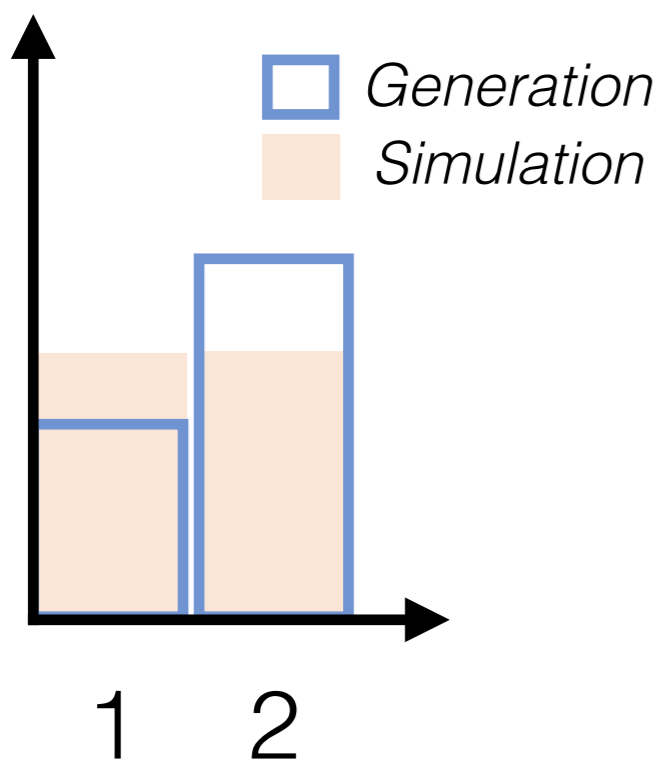


Measured	2	0%	50%
	1	100%	50%
		1	2
		Ideal	

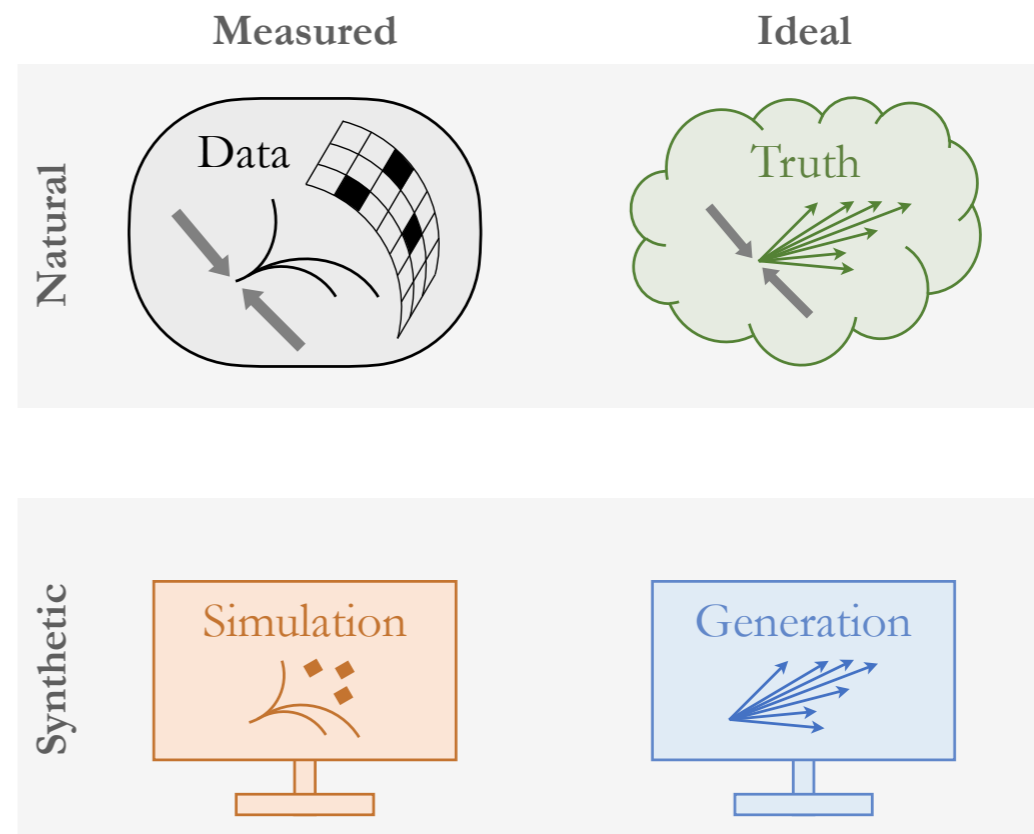


Unfold by iterating: OmniFold

After iteration 1

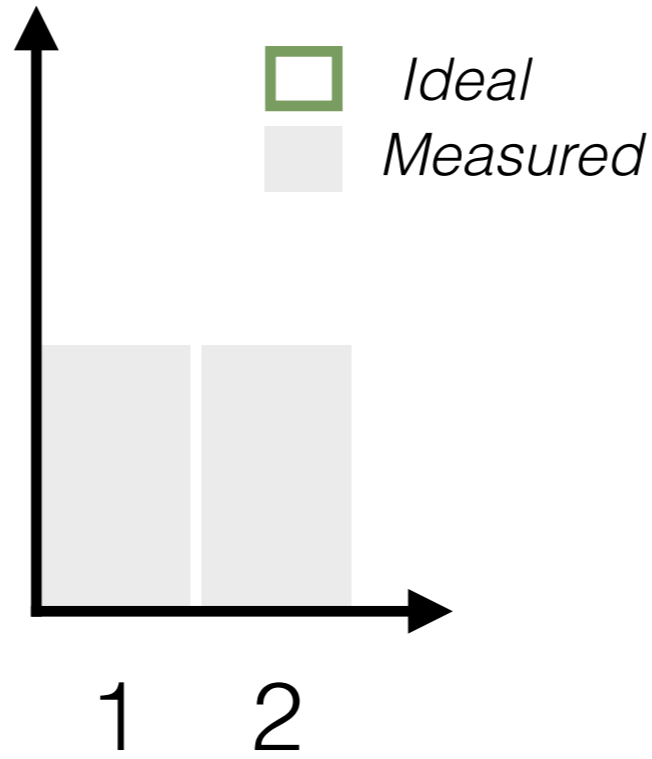
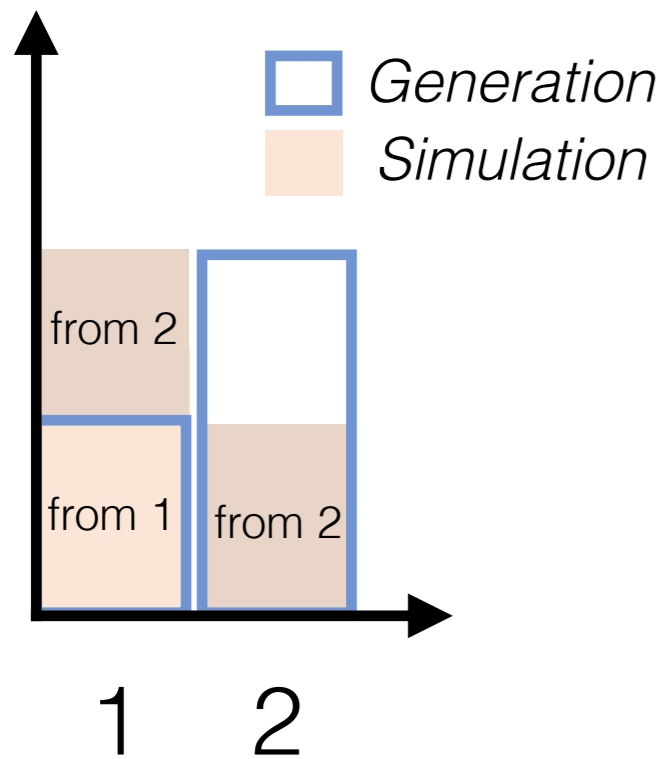


Measured	2	0%	50%
	1	100%	50%
		1	2
		Ideal	

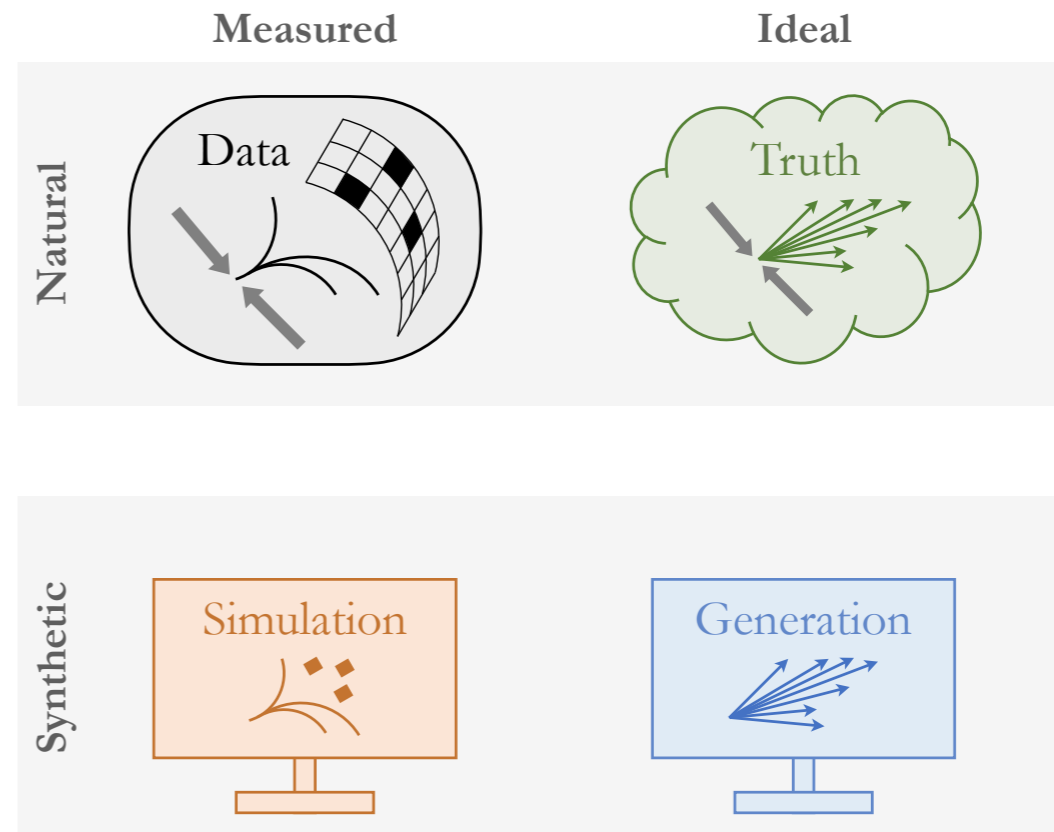


Unfold by iterating: OmniFold

After iteration 1

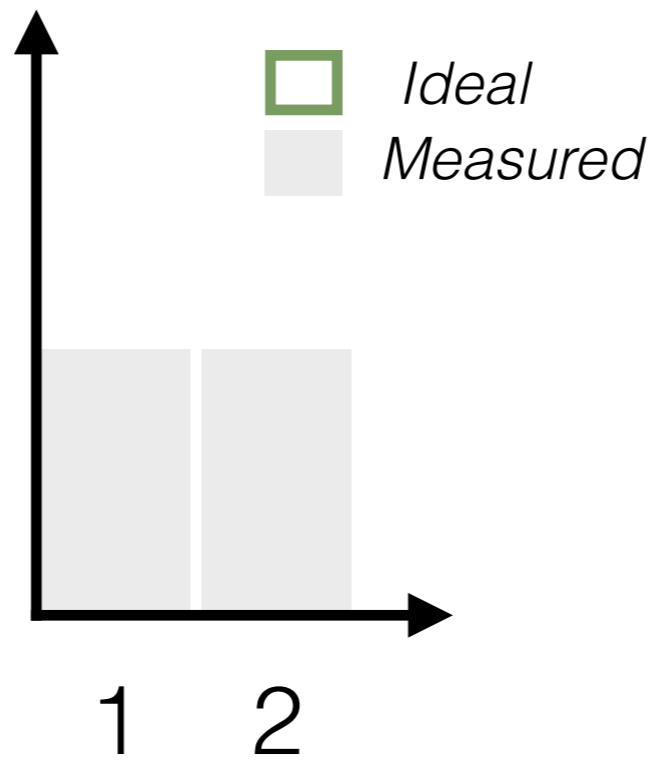
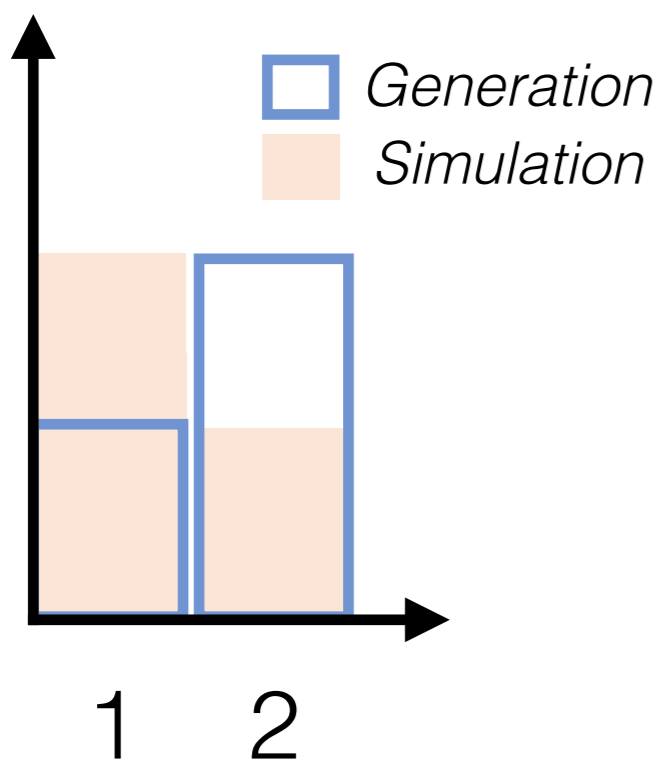


Measured	2	0%	50%
	1	100%	50%
		1	2
		Ideal	

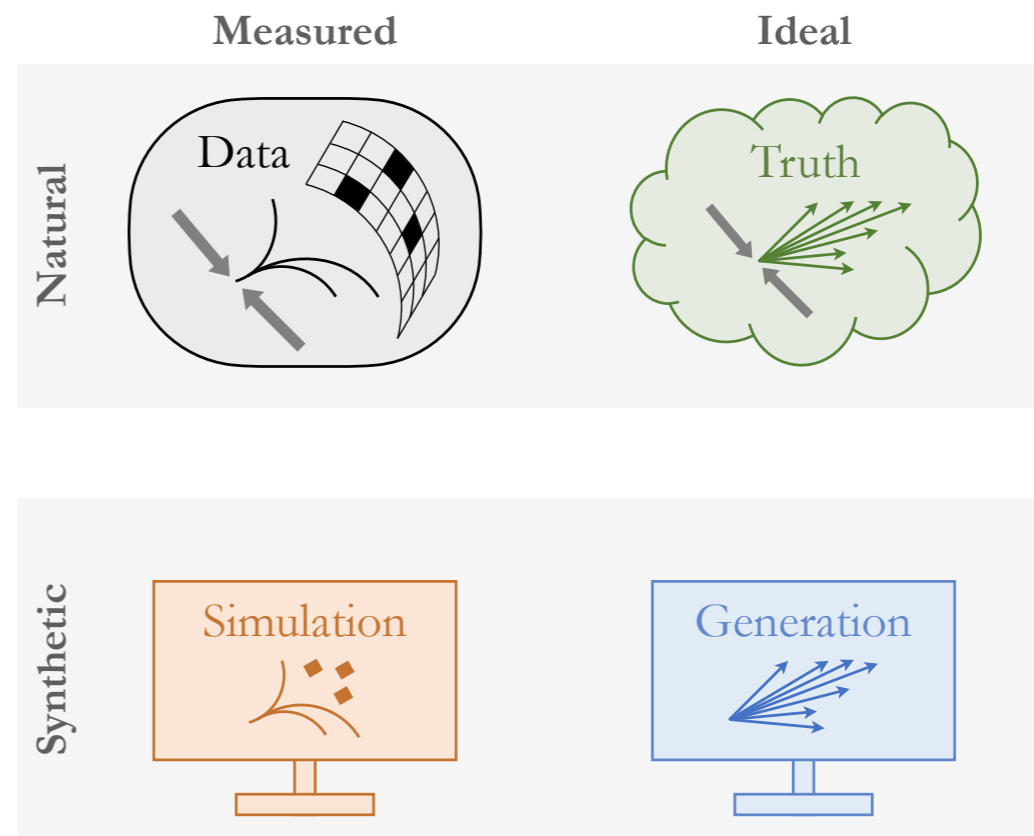


Unfold by iterating: OmniFold

After iteration 1

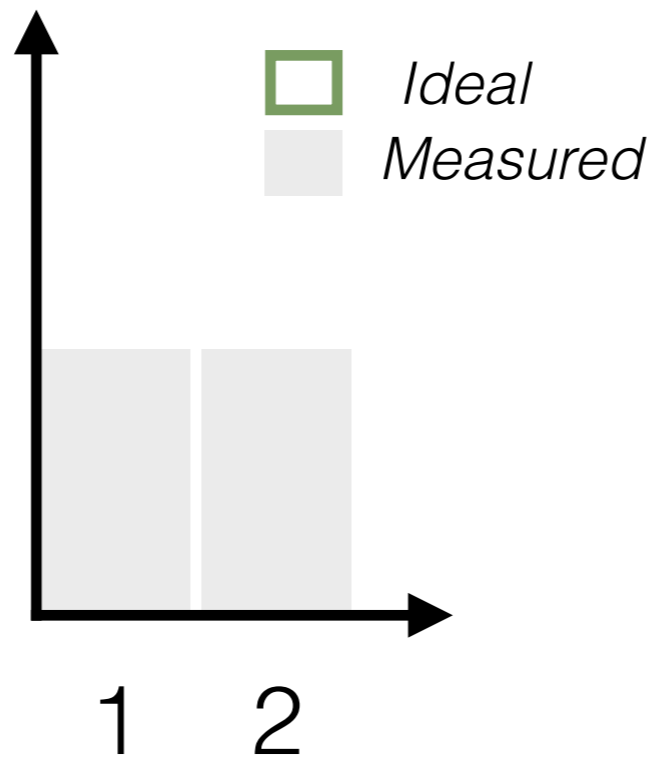
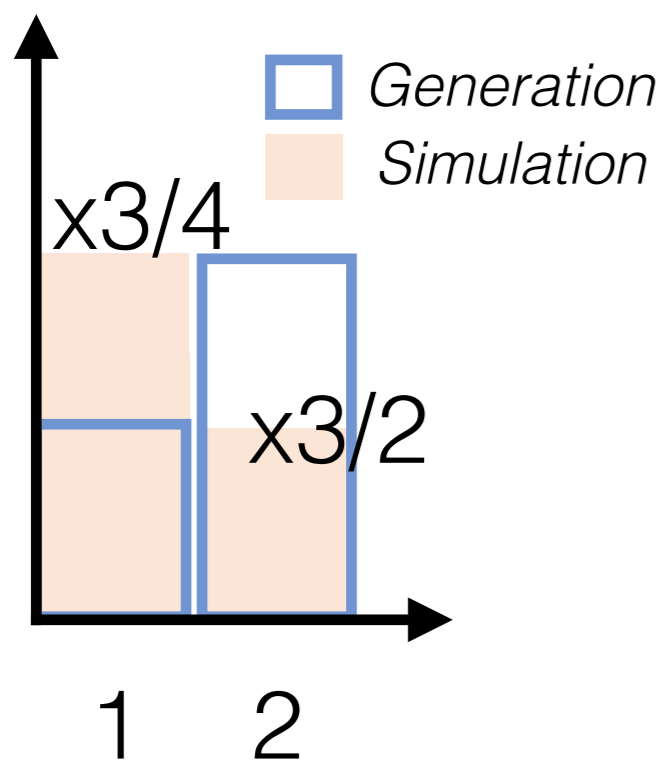


Measured	2	0%	50%
	1	100%	50%
		1	2
		Ideal	

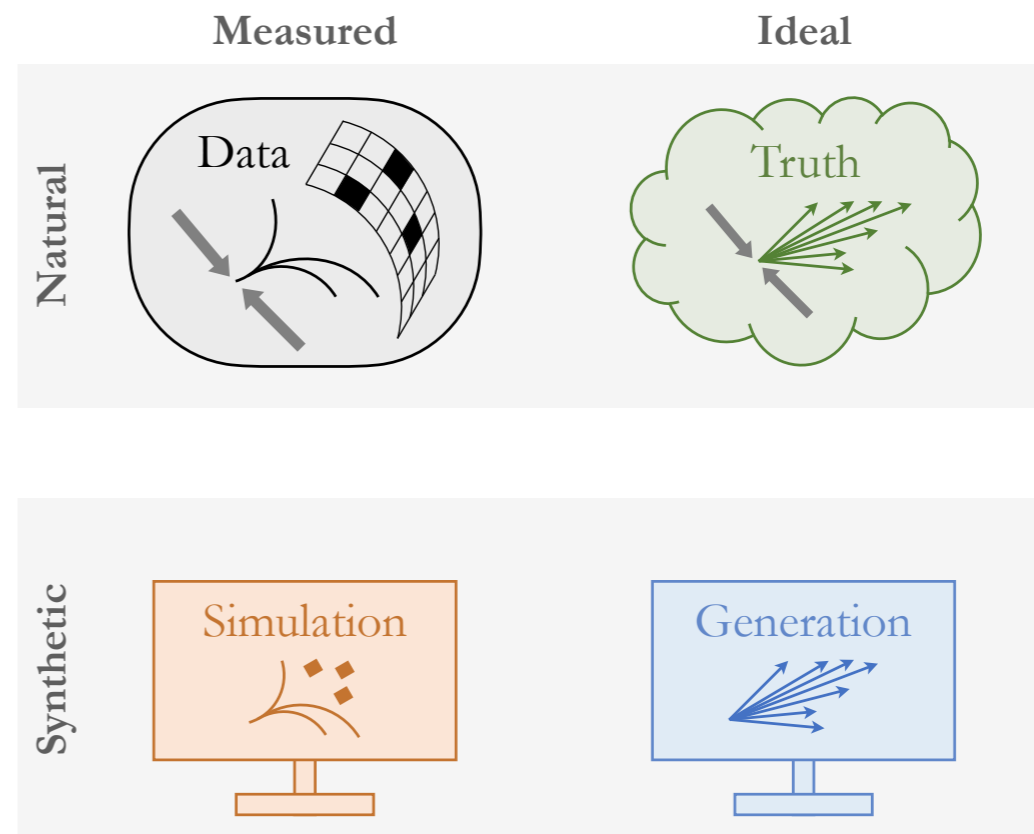


Unfold by iterating: OmniFold

After iteration 1

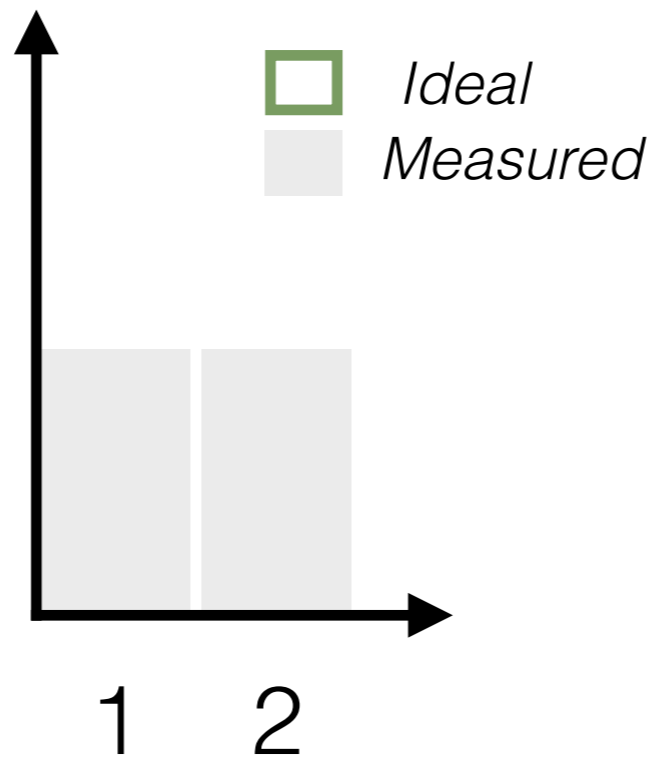
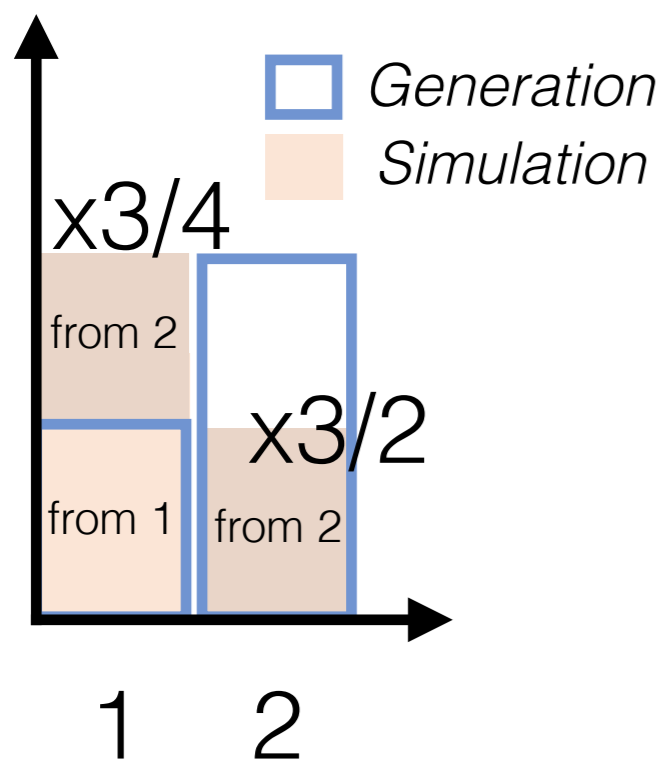


Measured	2	0%	50%
	1	100%	50%
		1	2
		Ideal	

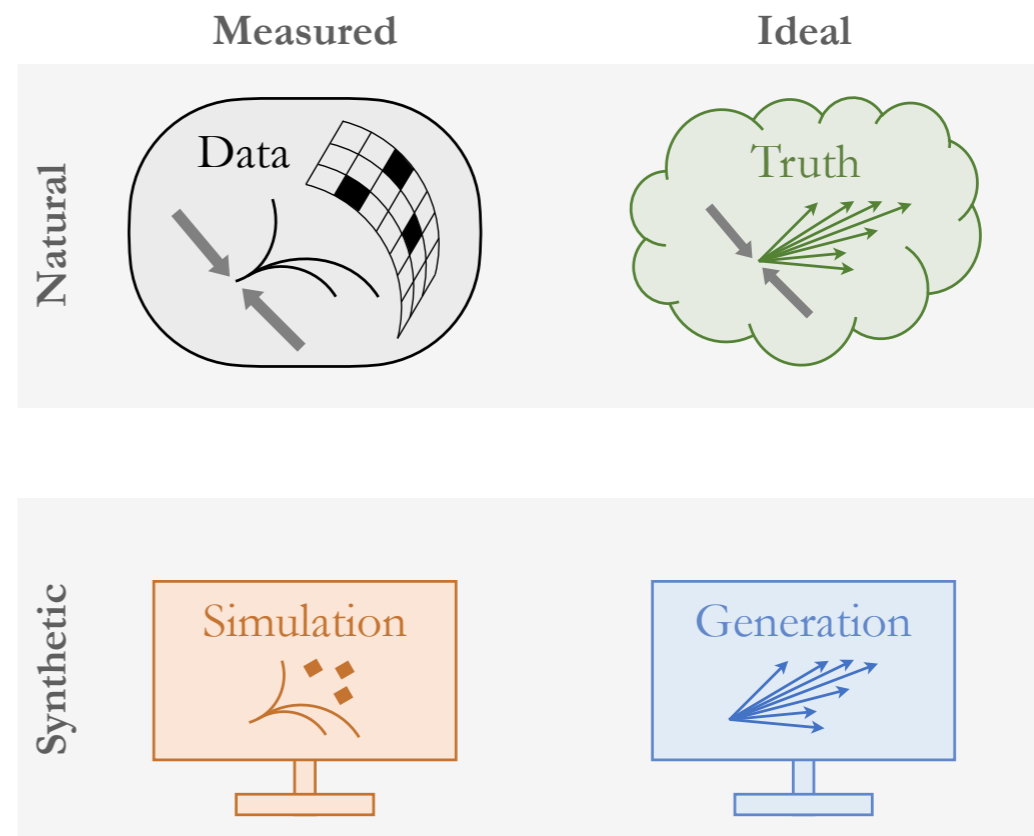


Unfold by iterating: OmniFold

After iteration 1

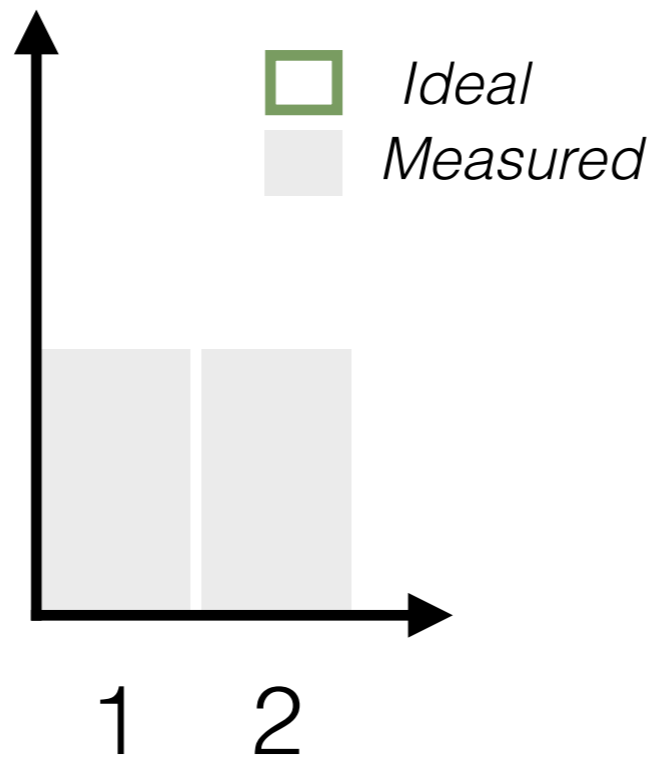
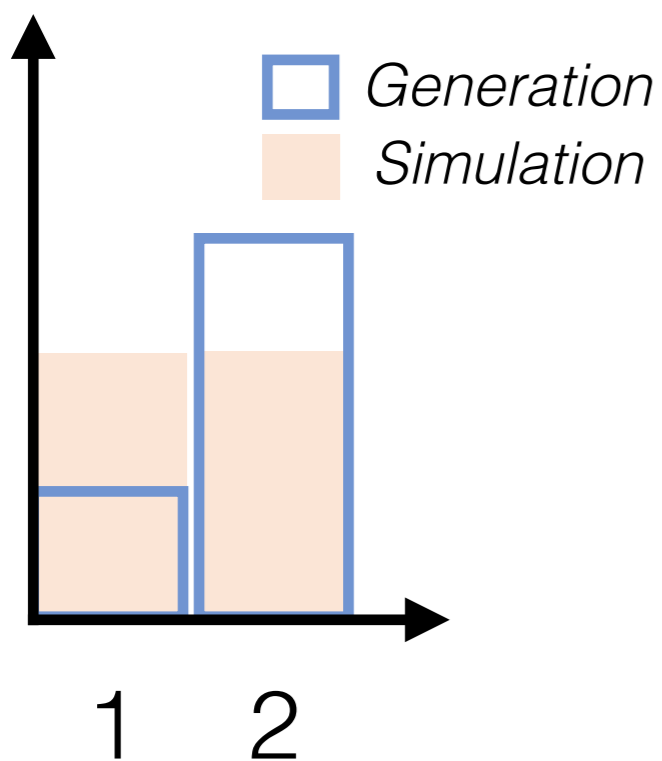


Measured	2	0%	50%
	1	100%	50%
		1	2
		Ideal	

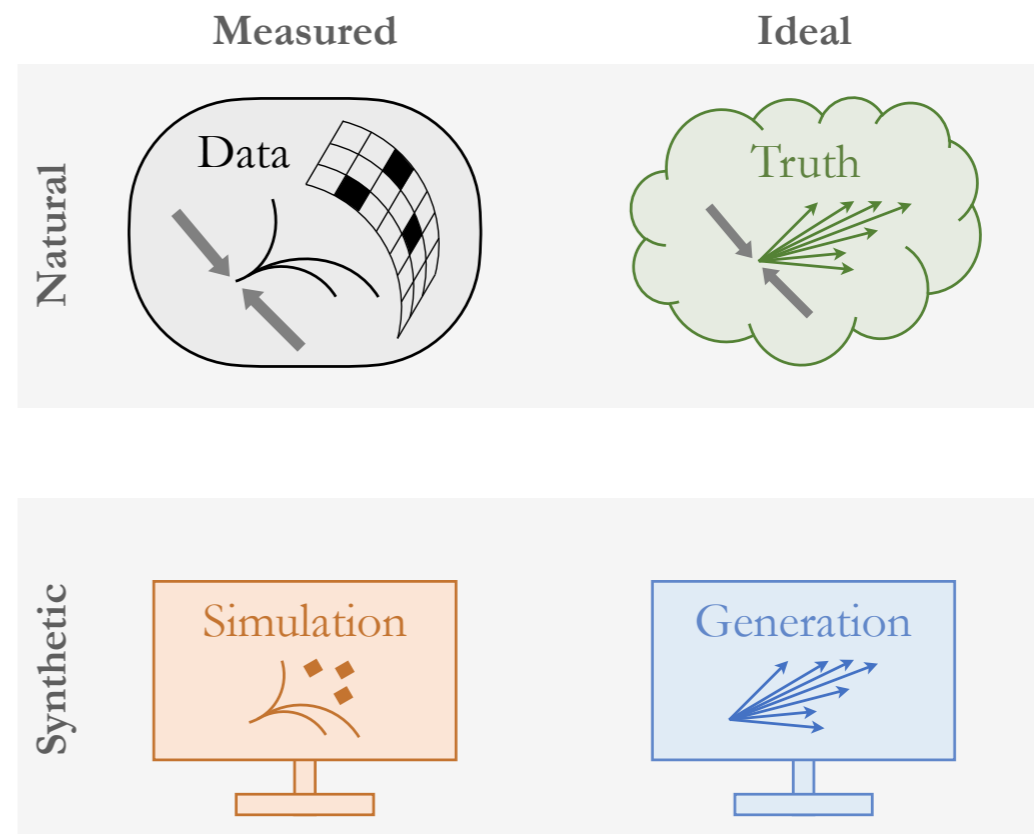


Unfold by iterating: OmniFold

After iteration 2

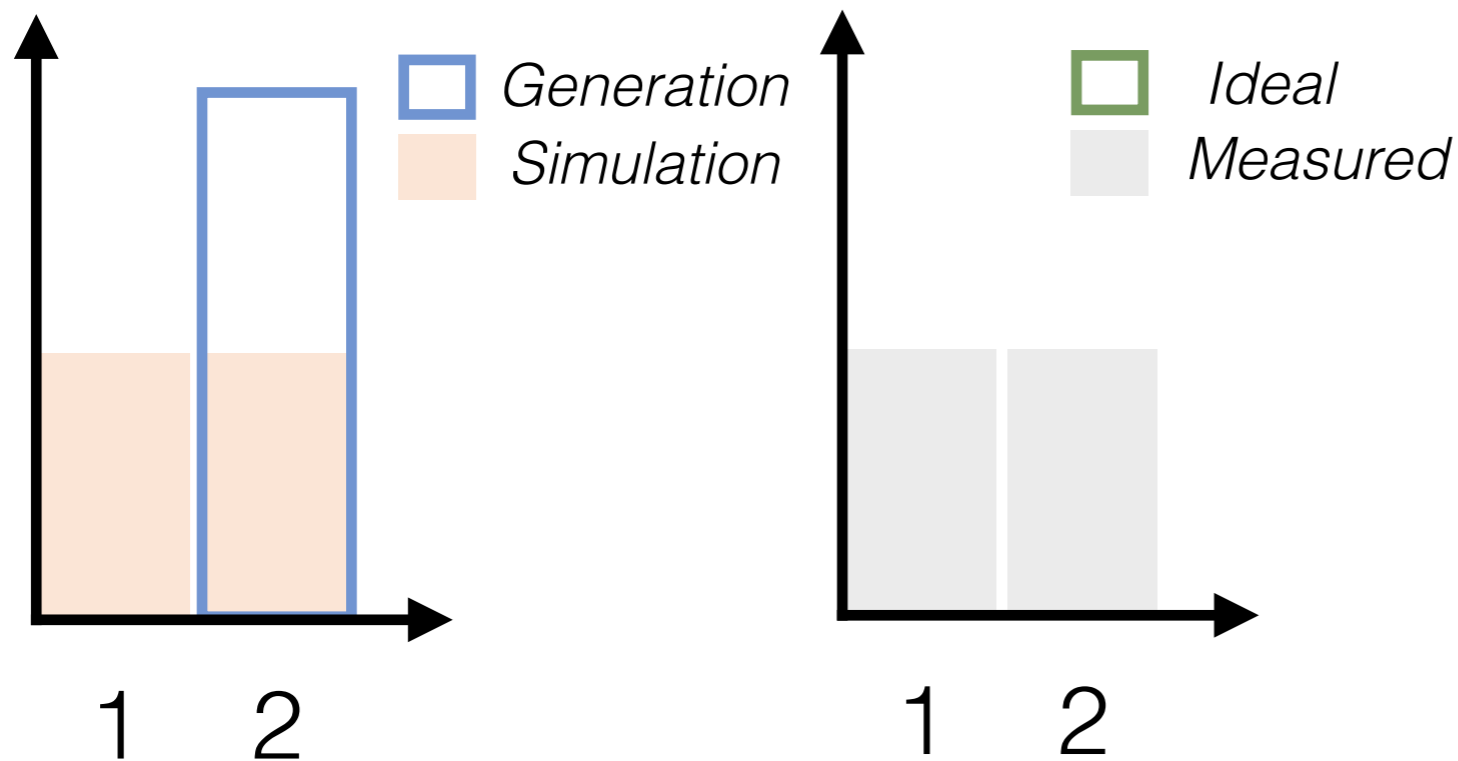


Measured	2	0%	50%
	1	100%	50%
		1	2
		Ideal	



Unfold by iterating: OmniFold

After iteration ∞



N.B. if you just apply $p(\text{ideal} | \text{measured})$, you would have gotten the wrong answer!

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

