Machine Learning for Collider Theory

— Snowmass Summer Study 2022 —

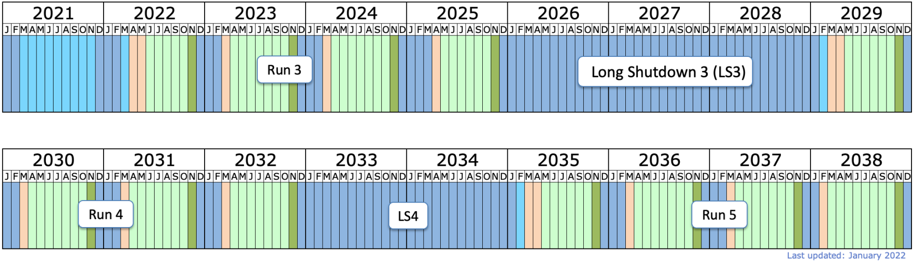
Claudius Krause

Rutgers, The State University of New Jersey

July 22, 2022

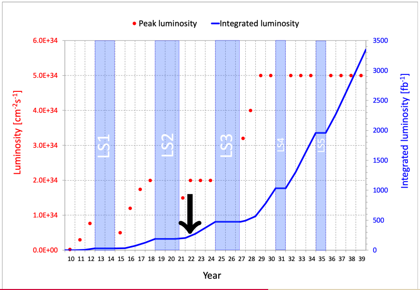
Based on the White Papers
Machine Learning and LHC Event Generation [2203.07460]
and
New directions for surrogate models and differentiable programming for HEP detector simulation [2203.08806]

We will have a lot more data in the near future.



Shutdown/Technical stop
Protons physics
Ions
Commissioning with beam
Hardware commissioning/magnet training

https://lhc-commissioning.web.cern.ch/schedule/LHC-long-term.htm



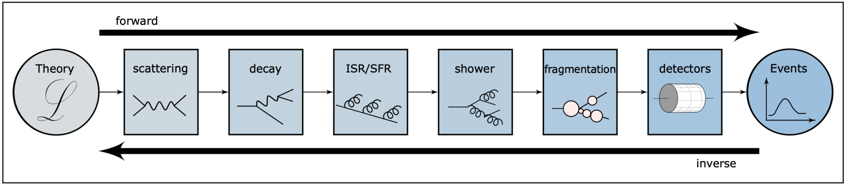
- We will have 20–25× more data.

$\Rightarrow$ We want to understand every aspect of it based on 1[st] principles!

We want to understand the LHC data based on 1^{st} principles.

What do we need to understand the data?

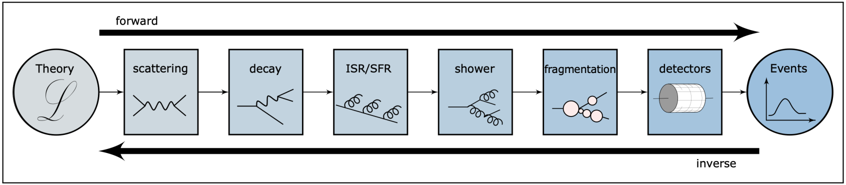
1. (a lot of) precise Simulations
2. optimized analyses for high-dimensional data



We want to understand the LHC data based on 1st principles.

What do we need to understand the data?

1. (a lot of) precise Simulations
2. optimized analyses for high-dimensional data

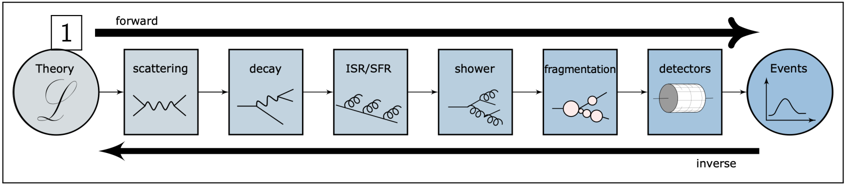


⇒ Machine Learning, as numerical tool, has a significant impact to every aspect of the simulation chain!

We want to understand the LHC data based on 1st principles.

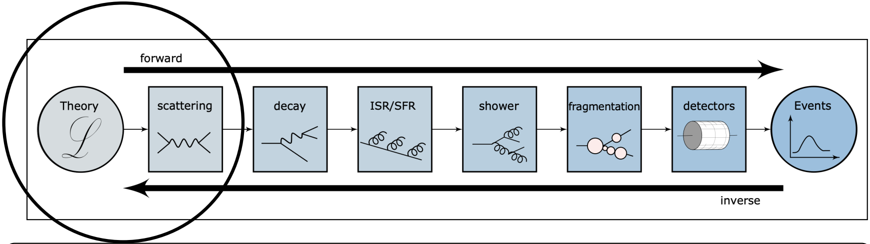
What do we need to understand the data?

1. (a lot of) precise Simulations
2. optimized analyses for high-dimensional data



⇒ Machine Learning, as numerical tool, has a significant impact to every aspect of the simulation chain!

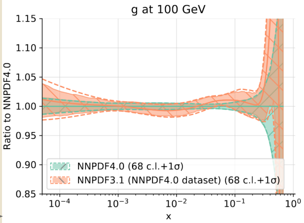
ML improves precise Simulations.



$$d\sigma \quad \sim \quad \text{pdf} \quad \times \quad \text{matrix element}^2 \quad \times \quad \text{phase space}$$

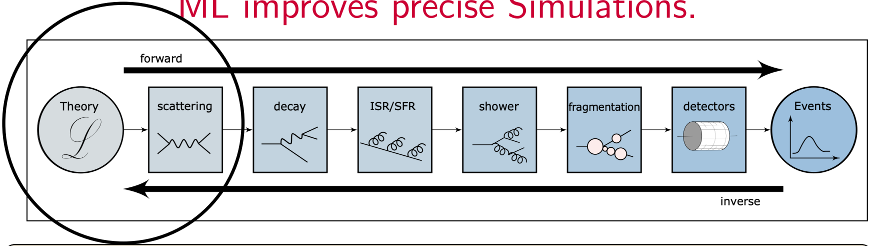
$\Rightarrow$ pdfs: ML reduces uncertainties.

- NNPDF uses NN for a long time (no parametric function assumed).
- Contemporary ML and hyperoptimizations reduced uncertainties from 3–5% to 1%.
- GAN-enhanced compression for delivery.



hep-ph/0204232, 1002.4407, 1410.8849, 2109.02671, 1907.05075, 2104.04535, . . . 2109.02653$\Rightarrow$

ML improves precise Simulations.



$$d\sigma \sim \text{pdf} \times \text{matrix element}^2 \times \text{phase space}$$

$\Rightarrow$ Amplitudes: Avoiding frequent calls to expensive matrix element.

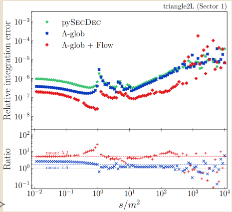
- as "simple" regression task,
- with uncertainties / boosted using a Bayesian NN,
- or using Catani-Seymour "basis" to reach per-mille level accuracy.

$\Rightarrow$ Loop integrals: increasing precision

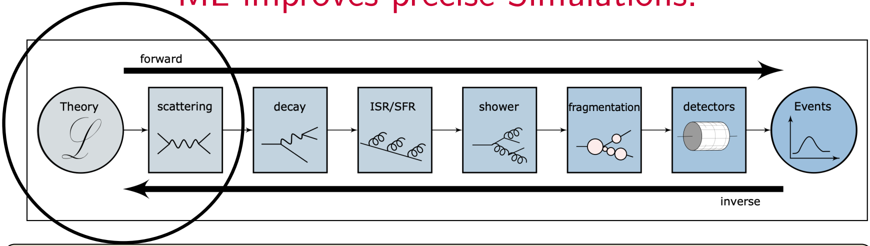
- NN-assisted contour deformation

2109.11964, 1912.11055, 2002.07516, 2106.09474, 2206.14831, 2107.06625, . . .

2112.09145 $\Rightarrow$



ML improves precise Simulations.



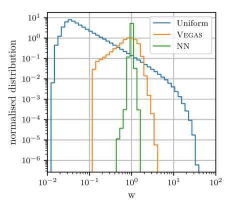
$$d\sigma \sim \text{pdf} \times \text{matrix element}^2 \times \text{phase space}$$

⇒ phase space: increase unweighting efficiency.

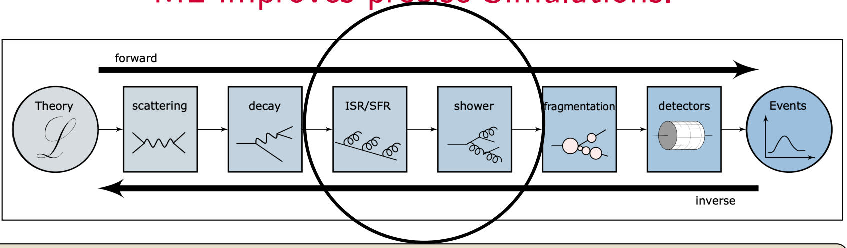
- improve Importance Sampling with normalizing flows
- learn channel weights in multi-channel integration
- learn distribution of events from event sample directly.

1707.00028, 1810.11509, 2001.05486, 2001.10028, 2009.07819, 2005.12719, 2011.13445, 1907.03764, 1903.02433, 1901.00875, 2102.11524, ...

2001.05478 ⇒



ML improves precise Simulations.



Semi-classical approximation good for small splitting angles.

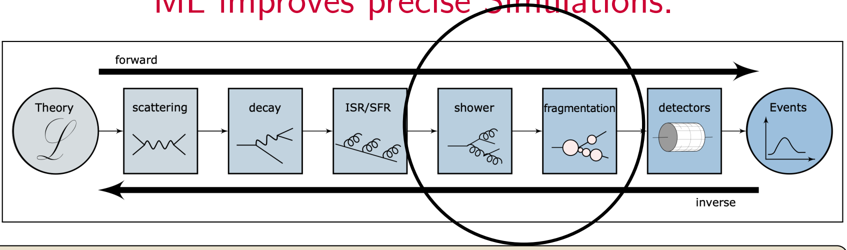
$\Rightarrow$ parton shower: improve over semi-classical approach

- splittings are iterative, can be learned by RNN;
- using ML-based inference to improve splitting kernels.

- Many body final states can be tackled by graphs or sets and learned directly.

1906.10137, 2012.06582, 1804.09720, 1808.07802, 1701.05927, 1807.03685, 2009.04842, 2109.15197, 2111.12849, 2012.09873

ML improves precise Simulations.

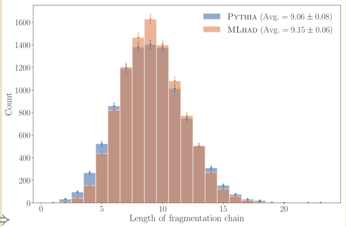


⇒ Fragmentation: Remove modeling bias.

- Same techniques as for pdfs.

⇒ Hadronization: better model non-perturbative effects.

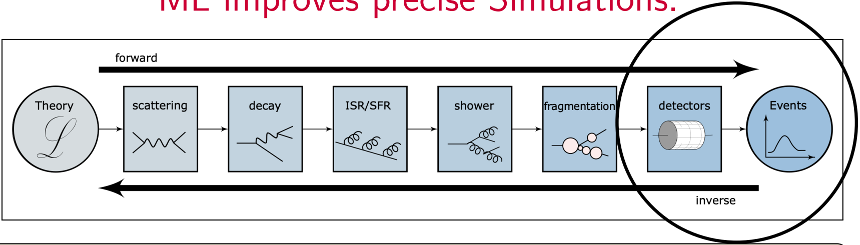
- Either improve existing clustering and Lund string model
- or use ML for more generic approach.



2105.08725, 1706.07049, 1807.03310, 2202.10779, ...

2203.04983 ⇒

ML improves precise Simulations.



⇒ Detector Simulation: Speed-up full GEANT4

- Indistinguishable showers $10^4\times$ faster.
- More ideas developed in "CaloChallenge 2022".

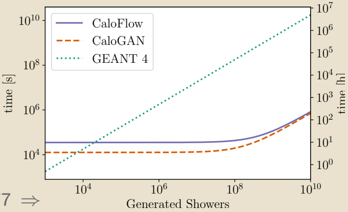
  https://calochallenge.github.io/homepage/

⇒ Trigger: more efficient storage and selection

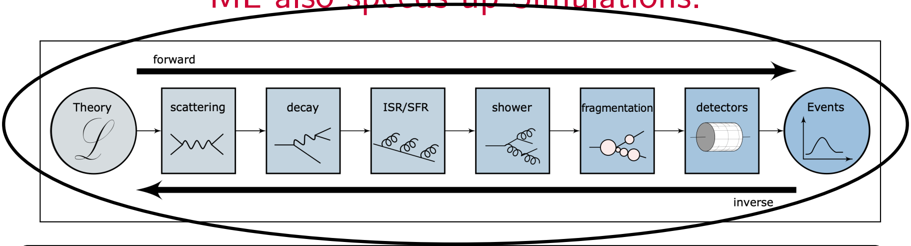
- Software for HLT, FPGAs for L1T

2109.02551, 1705.02355, 1712.10321, 1711.08813, 1802.03325, 1807.01954, 1912.06794, 2005.05334, 2102.12491, 2106.05285, PRL65.1321, 1712.07158, 1804.06913, 1903.10201, 2002.02534, 2104.03408, 2101.05108, 2110.13041, …

2110.11377 ⇒



ML also speeds up Simulations.



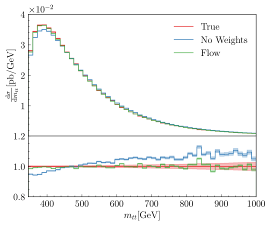
End-to-end ML-generators learn multiple steps at once. Advantages:

+ training on data combined with simulations,
+ post-processing of MC data for example to unweight events,
+ allows us to efficiently ship event samples,
+ provide datasets for phenomenological analyses,
+ enable inverted simulations,

⇒ Precise models with full control available!

2101.08944, 2110.13632, 1901.00875, 1901.05282, 1903.02433, 1912.02748, 2001.11103...

2011.13445 ⇒

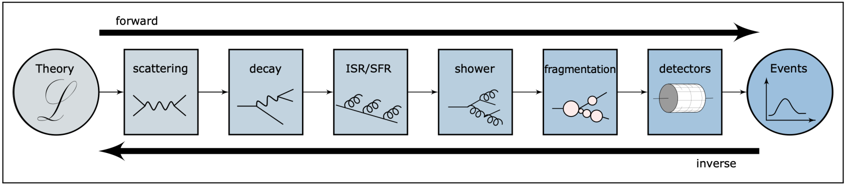


We want to understand the LHC data based on 1st principles.

What do we need to understand the data?

1. (a lot of) precise Simulations

2. optimized analyses for high-dimensional data



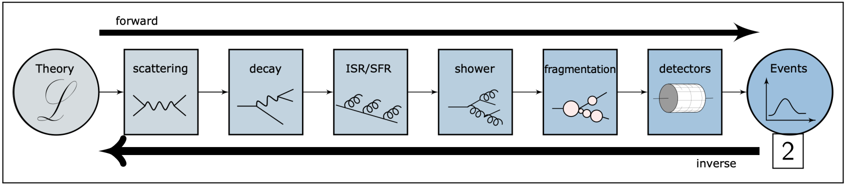
⇒ Machine Learning, as numerical tool, has a significant impact to every aspect of the simulation chain!

We want to understand the LHC data based on 1st principles.

What do we need to understand the data?

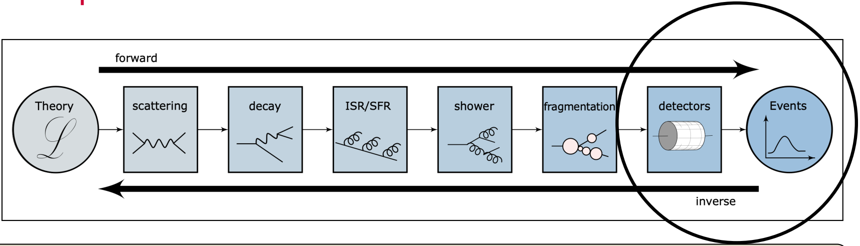
1. (a lot of) precise Simulations

2. <u>optimized analyses for high-dimensional data</u>



⇒ Machine Learning, as numerical tool, has a significant impact to every aspect of the simulation chain!

ML helps to invert the Simulation — for better inference.



⇒ Reconstruction in a busy detector

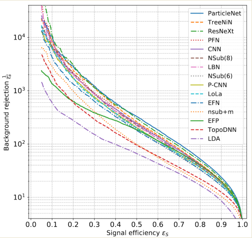
- going beyond traditional particle flow algorithms with GNNs
- improved tracking

⇒ Better particle identification

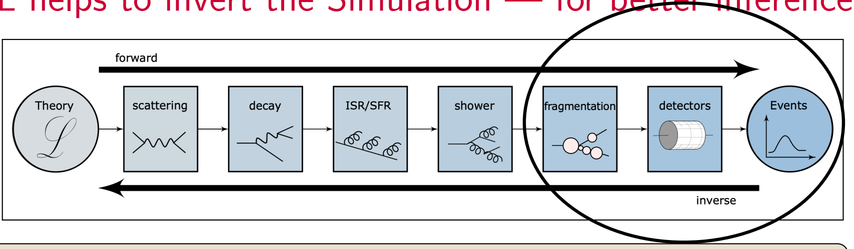
- e.g. Top Tagging Challenge

2003.08863, 2101.08578, 2106.01832,
2012.11944, 2012.04533, ...

1902.09914 ⇒

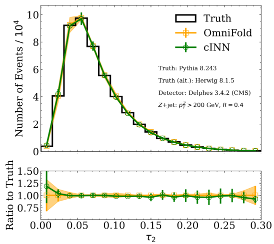


ML helps to invert the Simulation — for better inference.



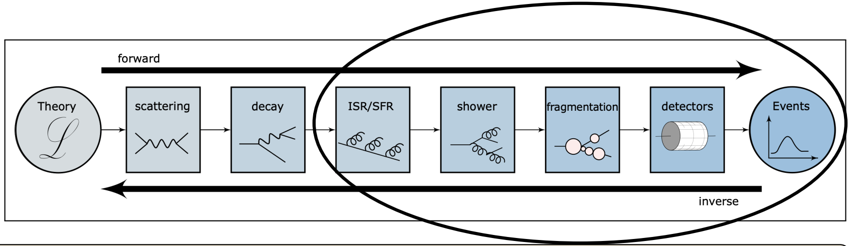
⇒ Unfolding of detector effects

- must be high-dimensional, unbinned, and statistically well-defined
- Classifier-based MC reweighting
- Conditional normalizing flow (cINNs)-based learn probability density per event



1806.00433, 2006.06685, 1911.09107, 2011.05836, 1912.00477, 2105.04448, 2105.09923, 2108.12376, . . . 2109.13243 ⇒

ML helps to invert the Simulation — for better inference.

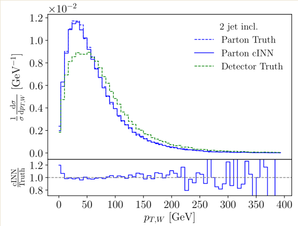


$\Rightarrow$ Inverting to parton level

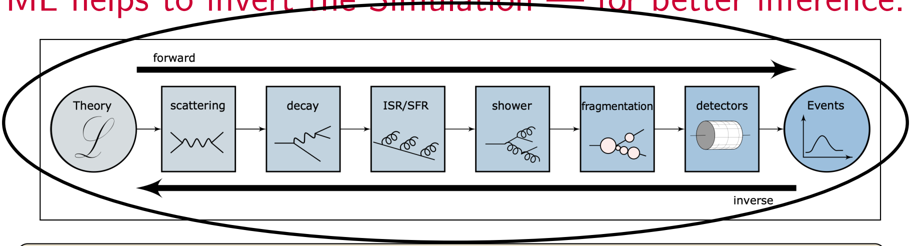
- Inversion of QCD radiation and heavy particle (t, W, Z, h) decays
- Uses similar techniques like unfolding (cINNs and Classifiers)

1806.00433, 2109.13243, 1911.09107, 2011.05836, 1912.00477, 2105.04448, 2105.09923, 2108.12376, . . .

2006.06685$\Rightarrow$



ML helps to invert the Simulation — for better inference.



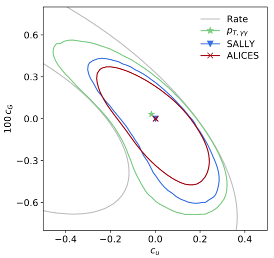
Simulation-based inference: inverting the full simulation chain

⇒ MadMiner

- can learn LL ratio or score by "mining" simulators.
- Having them differentiable would make it even more applicable.

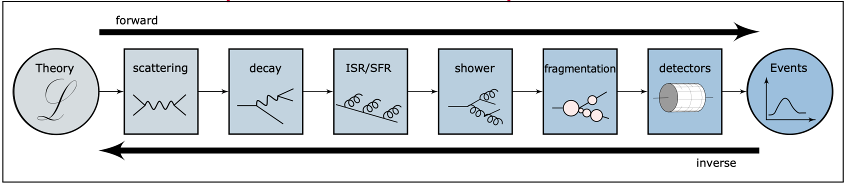
⇒ Matrix-Element Method (MEM)

- get LL from ME and unfolded events.
- More advanced (ML-based) unfolding yields better estimators.



1805.00013, 1805.00020, 1805.12244, 1808.00973, 2110.07635, 1506.02169, 1601.07913, . . . 1907.10621⇒

ML also improves model-independent searches.



Removing signal-model dependence to search for new physics:

⇒ Enhancing bump hunts

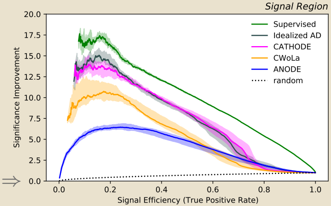
- methods now reaching results of idealized comparisons.
- lot's of active research in feature selection etc.

⇒ Anomaly detection

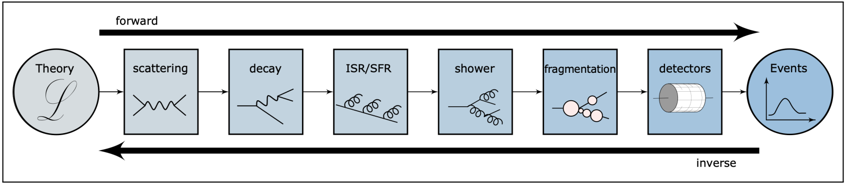
- searches for out-of-distribution events using various approaches
- see also "Dark Machines Anomaly Score Challenge"

1805.02664, 1902.02634, 1708.02949, 2001.04990, 2101.08320, 2105.14027, . . .

2109.00546 ⇒



The importance of ML for Collider Physics.



⇒ Modern ML is a new tool in our numerical toolbox — with applications to every step in the simulation/inference chain.

⇒ We've seen everything between "proof-of-concepts" to well established use cases.

⇒ There is an interesting interplay between HEP and the ML/AI community:

- Precise HEP simulations provide infinite, excellent training data for ML.
- HEP-specific application requirements (precision, symmetry, . . .) are different from industry applications (computer vision, etc.).