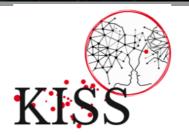
Painting pretty pictures of particles: Generative Al for detector simulation and new physics. seover

Gregor Kasieczka Email: gregor.kasieczka@uni-hamburg.de Twitter: @GregorKasieczka LPSC Grenoble Seminar 15.6.2023

CLUSTER OF EXCELLENCE

QUANTUM UNIVERSE









für Bildung

und Forschung

Bundesministerium



CENTER FOR DATA AND COMPUTING IN NATURAL SCIENCES

CDCS

Partnership of Universität Hamburg and DESY Please list in a concise way the three key developments expected for particle physics in the next decade

As an Al, I don't predict the future, but as of my last update in September 2021, the following are some key developments expected in the field of particle physics over the next decade based on the direction of ongoing research and proposed experiments:

S

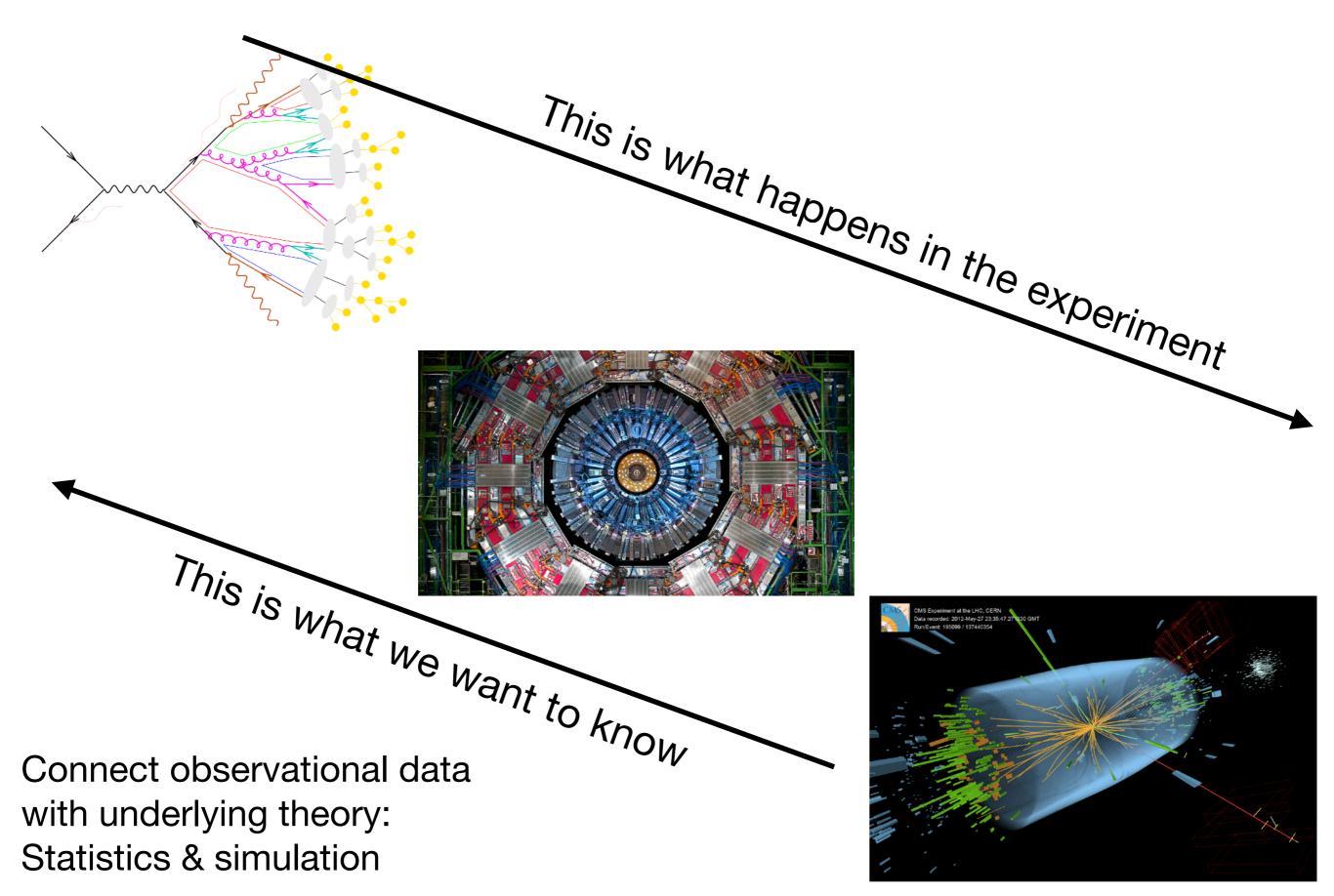
- 1. Completion of the High-Luminosity Large Hadron Collider (HL-LHC): This upgrade to CERN's Large Hadron Collider (LHC) will greatly increase the collider's luminosity, resulting in much more data for physicists to analyze. The HL-LHC could help to answer questions about the Higgs boson, dark matter, and supersymmetry. The project started in 2019 and is expected to be completed by the middle of this decade.
- 2. Progress in Neutrino Physics: Several experiments around the world, like the Deep Underground Neutrino Experiment (DUNE) in the U.S. and the Hyper-Kamiokande in Japan, aim to learn more about the properties of neutrinos, particularly their mass hierarchy and charge conjugation parity (CP) violation. If neutrinos are found to violate CP symmetry, it could provide a clue as to why the universe is made mostly of matter and not antimatter.
- 3. Advancements in Dark Matter Detection: Dark matter is one of the most pressing unsolved mysteries in physics. Experiments like the Xenon1T and its successor the Xenon-nT in Italy, and the Large Underground Xenon (LUX) and its upgrade, LUX-ZEPLIN (LZ) in the U.S., aim to directly detect dark matter particles. The next decade may see a breakthrough in our understanding of this elusive form of matter.

It's important to note that while these are major expected developments, the field of particle physics is vast and inherently unpredictable. Unexpected discoveries could lead to entirely new lines of inquiry.



a baroque oil painting of joseph fourier as DJ in a dance club

Experimental particle physics workflow



The rise of ML in particle physics

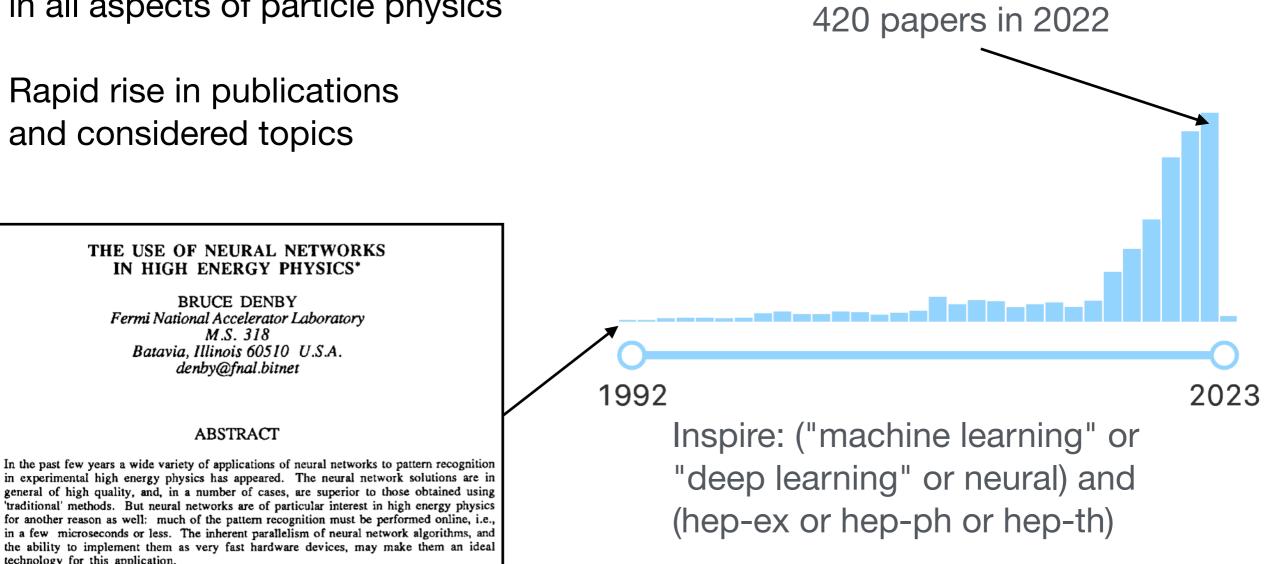
- Machine learning (ML) considered in all aspects of particle physics
- Rapid rise in publications and considered topics

BRUCE DENBY

M.S. 318 Batavia, Illinois 60510 U.S.A. denby@fnal.bitnet

ABSTRACT

technology for this application.



The rise of ML in particle physics

- Machine learning (ML) considered in all aspects of particle physics
- Rapid rise in publications and considered topics
- To maximise impact: Solid understanding of physics and machine learning techniques required
 - Domain knowledge of the physics data — e.g. symmetries
 - Phrase the actual physics task tractable to direct optimisation

See e.g. for a broader overview: <u>https://arxiv.org/abs/2112.03769</u>

Focus on LHC-style analysis but underlying ideas applicable in general

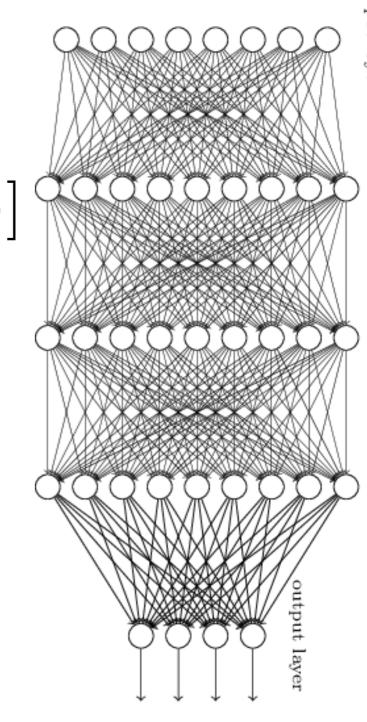
And particular to the second particle physics with the second particle physics. Various current and upcoming experiments are searching for signatures of new physics. Despite the variety of approaches and theoretical models tested in these experiments, what they all have in common is the very large volume of complex data that they produce. This data challenge calls for powerful statistical methods. Machine learning has been in use in high-energy particle physics for well over a decade, but the rise of deep learning in the early 2010s has yielded a qualitative shift in terms of the scot and ambition of research. These modern machine learning developments are the focus of the present Review, which discusses methods and applications for new physics searches in the com of terrestrial high-energy physics experiments, including the Large Hadron Collider, rare event searches and neutrino experiments. For several decades, the standard model (SM) of particle physics for new physics searches in the com of terrestrial high-energy physics experiments, including the Large Hadron Collider, rare event searches on an eutrino experiment. Authority is the tast tools will qualitative and process of nature beyond the SM (RSM). For example, the existence of draw matter (DM) and dark revers the tomps of some statistry to hypervariate analysis the estimation, variable is the targe of an ambition or synchice. In the Diverse of the nerespinent down and forces of nature beyothes Mindba the targe heat callenge is the site of the digene mark of the interesting method is allow for new analysis strategies independe the cistence of the generations of fermions, and the target need tools will qualitative the transmeters of the neurophilem by the search of first for new particles. This data to all light on all to the growing data challenge, there is the composing of high-energy physics (HEP) experiments are of the anteriations of the search of the anteriations of the search of the antenerging hysics (HEP) experiments, resulting the lea			REVIEWS
 Fundamental physics Georgia Karagiorgio ¹⁵⁸, Gregor Kasieczka²⁵⁸, Scott Kravitz ^{6,519}, Benjamin Nachman ⁶ and David Shiho ⁵²⁸ Abstract Compelling experimental evidence suggests the existence of new physics beyond the well-established and tested standard model of particle physics. Various current and upcoming experiments are searching for signatures of new physics. Despite the variety of approaches and theoretical models tested in these experiments, what they all have in common is the very large volume of complex data that they produce. This data challenge calls for powerful statistical methods. Machine learning has been in use in high-energy particle physics for well over a decar but the rise of deep learning in the early 2010s has yielded a qualitative shift in terms of the scop and ambition of research. These modern machine learning developments are the focus of the present Review, which discusses methods and applications for new physics searches in the com of terrestrial high-energy physics experiments, including the Large Hadron Collider, rare event searches and neutrino experiments. 			Check for updates
 and David Shino⁵⁵⁶³ Abstract Compelling experimental evidence suggests the existence of new physics beyond the well-established and tested standard model of particle physics. Vespite the variety of approaches and theoretical models tested in these experiments, what they all have in common is the very large volume of complex data that they produce. This data challenge calls for powerful statistical methods. Machine learning has been in use in high-energy particle physics for well over a decar but the rise of deep learning in the early 2010s has yielded a qualitative shift in terms of the scop and ambition of research. These modern machine learning developments are the focus of the present Review, which discusses methods and applications for new physics searches in the com of terrestrial high-energy physics experiments, including the Large Hadron Collider, rare event searches and neutrino experiments. For several decades, the standard model (SM) of particle physics has provided a clear theoretical guide to experiments, resulting in an extensive search programe that culminated with the discovery of the Higgs boson¹¹. Although the SM is now complete, there are learts of forts for new particles in analysis due to forse of nature beyond the SM (BSM). For example, the existence of thexe baser hefforts for new particles in the sensitivity of hypervariate analysis (derst) estimation, variable and forces of nature beyond the SM (BSM). For example, the existence of thexe observations are explained by the SM and distionally, assthetic problems plaque the SM, induce and the distingal base of the enurchaling development are explained weak-seale mass of the Higgs boson¹¹. Additionally, assthetic problems plaque the SM, induce the groveing atta volume, a red maintenes of the neuron dipole moment? Current and include the development are future high-energy physics (HEP) experimenta for the revise of the experiments may observe. HEP experiments is utowersing or is unwersing or is u			
 well-established and tested standard model of particle physics. Various current and upcoming experiments are searching for signatures of new physics. Despite the variety of approaches and theoretical models tested in these experiments, what they all have in common is the very large volume of complex data that they produce. This data challenge calls for powerful statistical methods. Machine learning has been in use in high-energy particle physics for well over a decar but the rise of deep learning in the early 2010s has yielded a qualitative shift in terms of the scop and ambition of research. These modern machine learning developments are the focus of the present Review, which discusses methods and applications for new physics searches in the cont of terrestrial high-energy physics experiments, including the Large Hadron Collider, rare event searches and neutrino experiments. For several decades, the standard model (SM) of particle physics has provided a clear theoretical guide to experiments, resulting in an extrems versity search pregrament, resulting in an extrems versity searches and heutrino experiments. For several decades, the standard model (SM) of particle physics has provided a clear theoretical guide to experimental estimation for analysing large amounts of data in many dimense to figure search efforts for new particles and forces of nature beyond the SM (ISN). For example, there are the asso for neutrinos⁴⁴ and the sensitivity to 'hypervariate analysis' whe tentre phase space of available experimental exit tools will guiltant terms. Multivariate analysis has a loallow for new analysis has a call acluber of the existence of the experiments, resultated by the SM. Additionally, 'asthetic problems plague the SM, indust, guident the growing data values, a relet mass of the reutron dipole moment'. Current and the <i>Existence of three generations</i> of fereminons, and the achilenge the reguestion to the acevelopming the state that were platforms' and the the potential			³, Scott Kravitz⊚⁵≈, Benjamin Nachman⊚ ^{5,4} 5
 ticle physics has provided a clear theoretical guide to experimental observations and a created pro- grammethat culminated with the discovery of the Higgs boson^{1,2}. Although the SM is now complete, there are key experimental observations that compel the com- munity to expand the search efforts for new particles and forces of nature beyond the SM (ISSM). For example, the existence of dark matter (DM) and dark energy pie, the existence of dark matter (DM) and dark energy is well established², as are the mass of neutrinos¹⁵ and the biopro-antilaryon asymmetry in the Universe¹⁰. and forces on foature beyond the SM (ISSM). For example, the existence of dark matter (DM) and dark energy is well established², as are the mass of neutrinos¹⁵ and the biopro-antilaryon asymmetry in the Universe¹⁰. and so on). In tandem with the growing data volume, a rel is Ottion, lawrene key Nitoral thorough the existence of three generations of fermions, and accurate data processing for high-throughput a mental challenges by creating new particles with ensite and accurate data processing for high-throughput a ration to the optimit of other new particles in the constructive of who ther new particles in to can Astronomy, siderable data challenges. New particle interactions of research, 1 		well-established and tested standard model of p experiments are searching for signatures of new theoretical models tested in these experiments, volume of complex data that they produce. This methods. Machine learning has been in use in hig but the rise of deep learning in the early 2010s ha and ambition of research. These modern machin present Review, which discusses methods and ag of terrestrial high-energy physics experiments, in	article physics. Various current and upcoming physics. Despite the variety of approaches and what they all have in common is the very large data challenge calls for powerful statistical gh-energy particle physics for well over a decade, as yielded a qualitative shift in terms of the scope e learning developments are the focus of the polications for new physics searches in the contex
Image: Complexe is a complexe in the complexe is a complexe in the complexe is a complexe is complexe is a complexe is a complexe is a complexe is	¹ Department of Physics, Columbia University, New York, NY, USA Sinstant University Inneuroparation Experimental physics, Hornburg, Cermany, Hornburg, Cermany, Benchely, National Laboratory, Benchely, National Laboratory, Benchely, National Laboratory, Benchely, National Laboratory, Benchely, National Laboratory, Benchely, National Jacob Octower, Columbia Physics and Astronomy, Nutgers University, Prestamor, U. USA "email: georgical meria.commol.edus group: chastercolor But anti-handrug de subvoixed Bild part: physics and Astronomedia Bild anti-handrug de subvoixed Bild anti-handrug de subvoixed Bild part.physics.anti-handrug de subvoixed Bild part.physics.anti-handrug de subvoixed Bild part.physics.anti-handrugers.anti-	ticle physics has provided a clear theoretical guide to experiments, resulting in an extensive search pro- gramme that culminated with the discovery of the Higgs boson ¹² . Although the SM is now complete, there are key experimental observations that complet the com- munity to expand the search efforts for new particles and forces of nature beyond the SM (RSM). For exam- ple, the existence of dark matter (DM) and dark energy is well established ² , as are the mass of neutrinos ²⁴ and the baryon-antibaryon asymmetry in the Universe ¹⁴ - yet none of these observations are explained by the SM. Additionally, aschetcir problems plague the SM, indud- ing the unexplained weak-scale mass of the Higgs boson, the existence of three generations of fermions, and the minuteness of the neutron dipole moment ⁷ . Current and near future high- energy physics (HEP) experiments have the potential to shed light on all of these fund- mental challenges by creating new particles in the labo- ratory, or by observing interactions of new particles. This great potential for discovery comes with con- siderable data challenges. New particle interactions are expected to be rare, and their signature could be only subty different from the SM. This means that researchers must collect and sift through an immense amount of complex data to isolate potential RSM physics. Machine learning (ML) offers a powerful solution to this chal-	tens of thousands of tunable parameters) are well suiter for analysing large amounts of data in many dimension to find subtle patterns. Multivariate analysis has beer commonplace in HEB for decades (for example, the TMVA 'toolkit'), but the latest tools will qualitatively extend the sensitivity to 'hypervariate analysis' whereby the entire phase space of available experimental infor mation can be analysed holistically. These new tool also allow for new analysis strategies independent o the dimensionality (density estimation, variable-lengti inputs and so on). In tandem with the growing data volume, a related challenge is the increasing need for efficient (in term of computational time, power and resource utilization and accurate data processing for high-throughput appli- cations. Efforts to that end include the developmen and acceleration of deep learning-based processing algorithms on power-efficient hardware platforms'. In addition to the growing data challenge, there is als the compounding challenge of simulating expectation or what coperiments may observe. HEP experiments rel heavily on simulations for all aspects of research, from coperiments design all the way to data malaysis. Buil on a thorough understanding of the SM and the funda- mental laws of nature. It is therefore often necessars to combine simulations with information directly from an approximation to nature. It is therefore often necessars to combine simulations with information directly from
	https://doi.org/10.1038/ s42254-022-00455-1		ML models must be robust against inaccuracies and b- able to integrate uncertainties.

Micro-Intro: Machine Learning

- Rephrase task as a minimisation problem..
- ..and "simply" solve:

$$\theta^* = \operatorname{argmin}_{\theta} \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})} \left[\mathcal{L}(f_{\theta}(\mathbf{x}), \mathbf{x}) \right]^{\theta}$$

- Modern ML: function f is a deep neural network & minimisation carried out via gradient descent
- Devil in the details:
 - How to map physics objective to loss function L
 - How to structure f to make maximum use of physics knowledge
 - How learn in a robust way from minimum amount of data



hidden layer

hidden layer 2

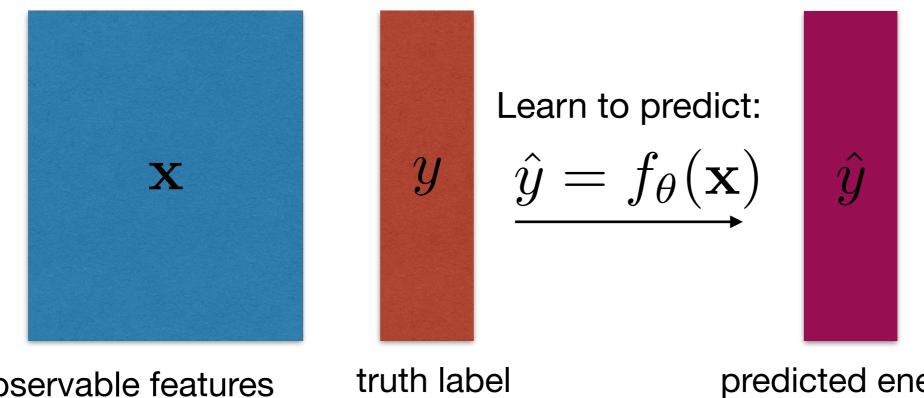
hidden layer 3

Loss function: Supervised

Supervised Learning:

Attempt to infer some target (truth label): classification, regression (often also clustering/inference)

Use training data with known labels (often from Monte Carlo simulation)



observable features such as kinematics, tracks,...

(e.g. true energy)

predicted energy

Regression: Minimize mean squared error:

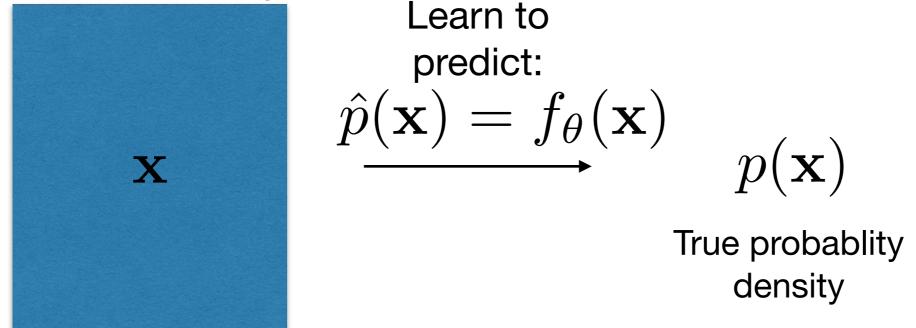
 $\mathcal{L} = (y - \hat{y})^2$

Loss function: Unsupervised

Unsupervised Learning:

No target, learn the probability distribution (directly from data)

Can use for sampling, anomaly detection, unfolding, ...



Distribution learning: Maximise likelihood (minimize log-likelihood): (either directly or with approximations)

$$\mathcal{L} = -\log\left(\hat{p}(\mathbf{x})\right)$$

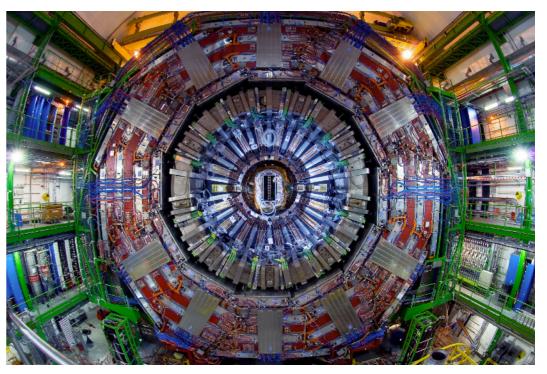
Fast simulation

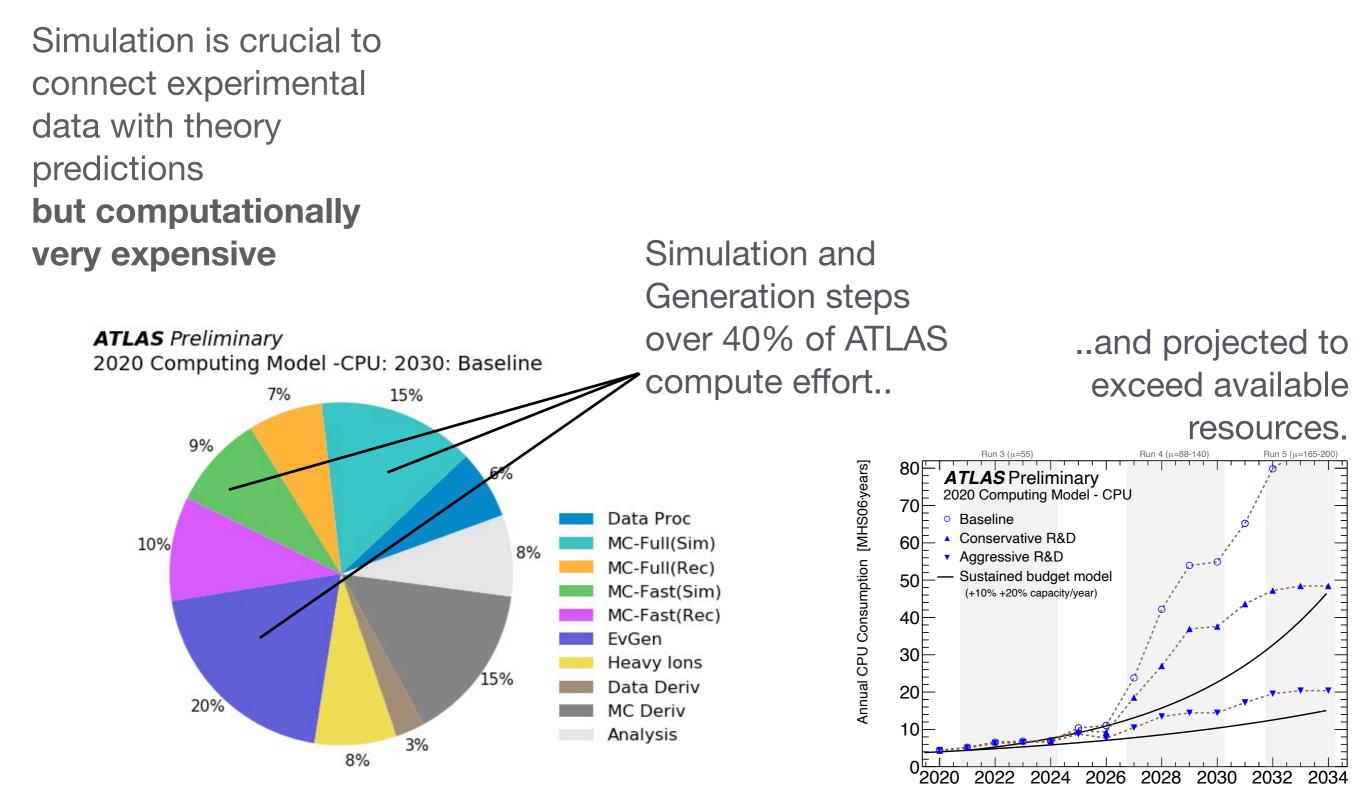


Simulation is crucial to connect experimental data with theory predictions

 $\begin{aligned} \mathcal{I} &= -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} \\ &+ i \mathcal{F} \mathcal{D} \mathcal{V} + h.c. \\ &+ \mathcal{V}_i \mathcal{Y}_j \mathcal{V}_j + h.c. \end{aligned}$ $+\left|\mathcal{D}_{M}\varphi\right|^{2}-\bigvee(\phi)$

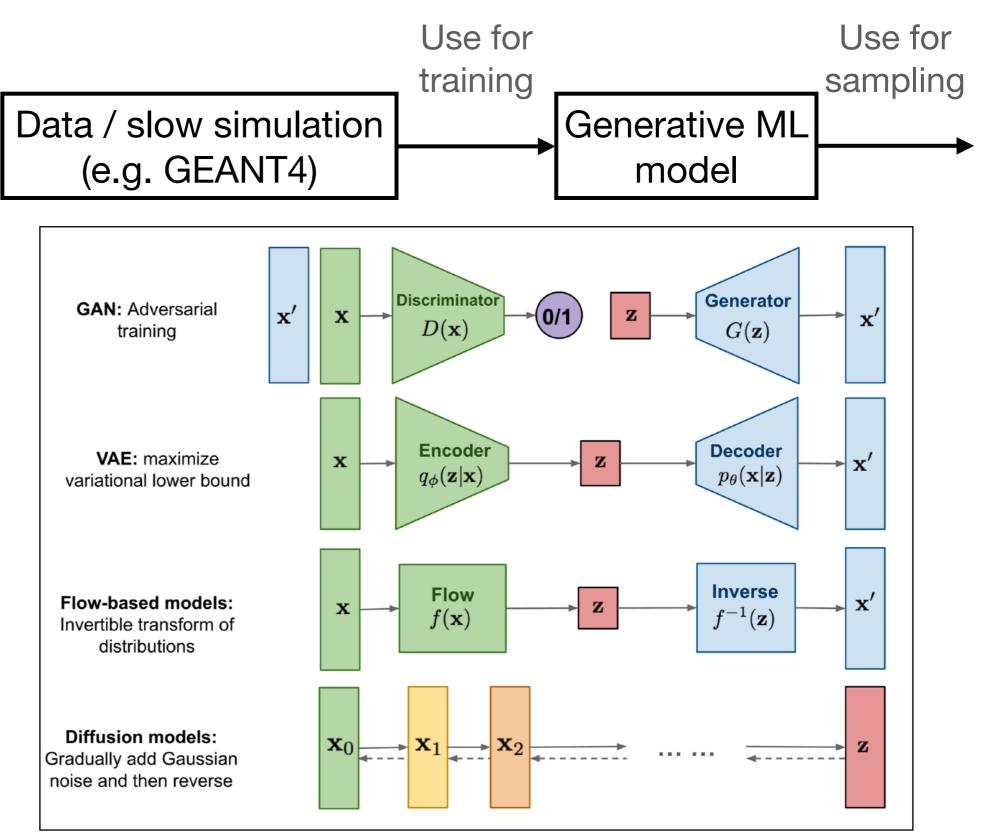
Complex chain of simulations

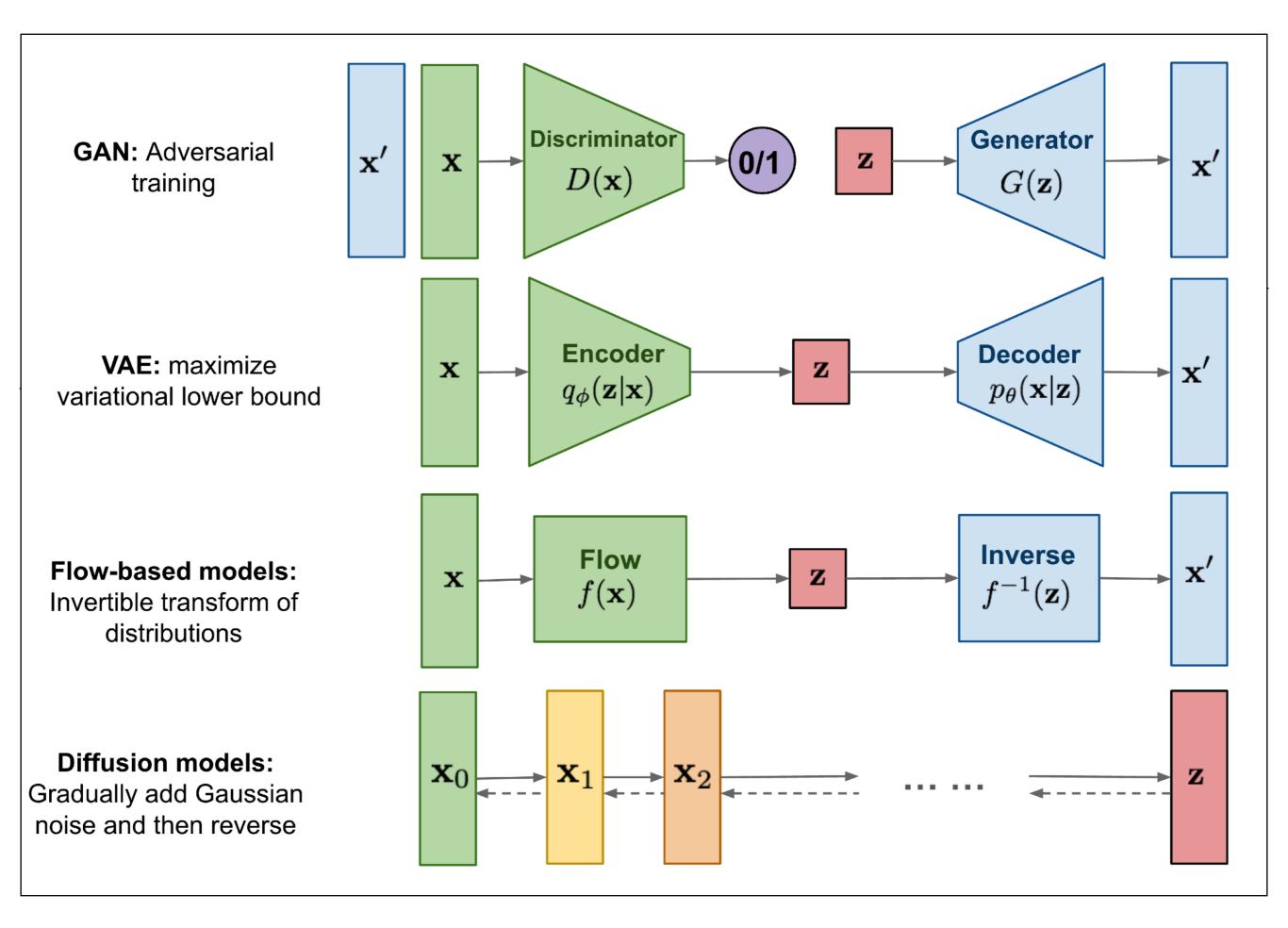




Simulation is crucial to connect experimental data with theory predictions **but computationally very expensive**

Use generative models trained on initial data to augment statistics





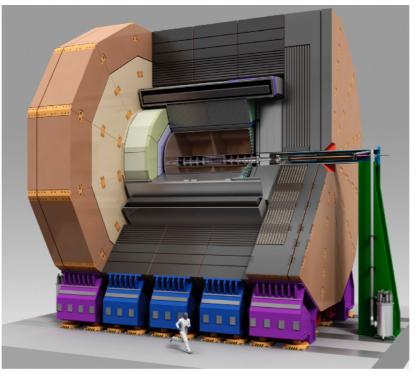
Simulation is crucial to connect experimental data with theory predictions **but computationally very expensive**

Use generative models trained on initial data to augment statistics

Goal: simulation of particle showers in complex calorimeters like for the ILD or the CMS High Granularity Calorimeter (HGCal)



CALICE AHCal testbeam



ILD Detector

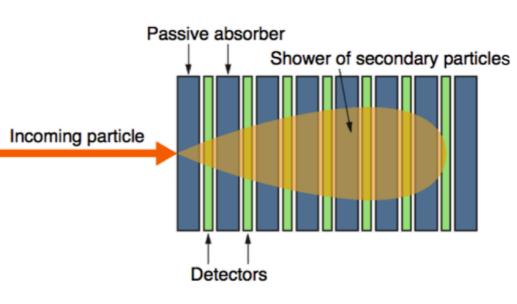
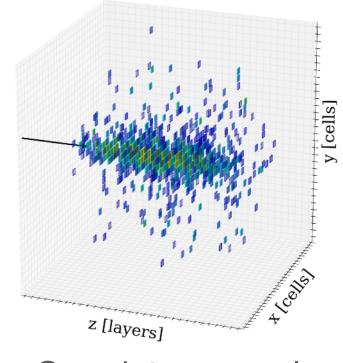


Illustration of particle shower in a sampling calorimeter



One data example

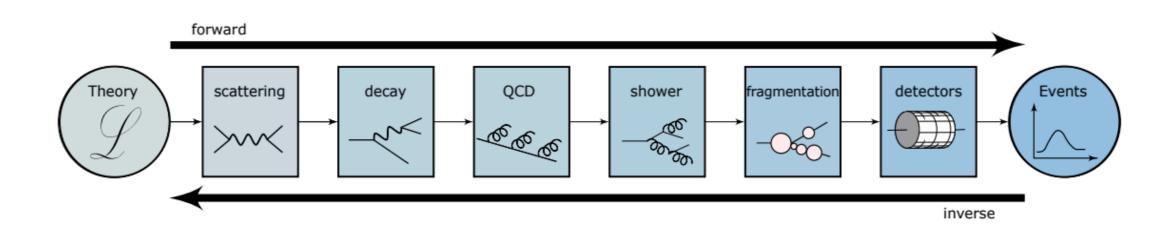
Aside

Large effort also exists to learn surrogate models of theory simulations // event generation

Similar techniques are used, but different key challenges

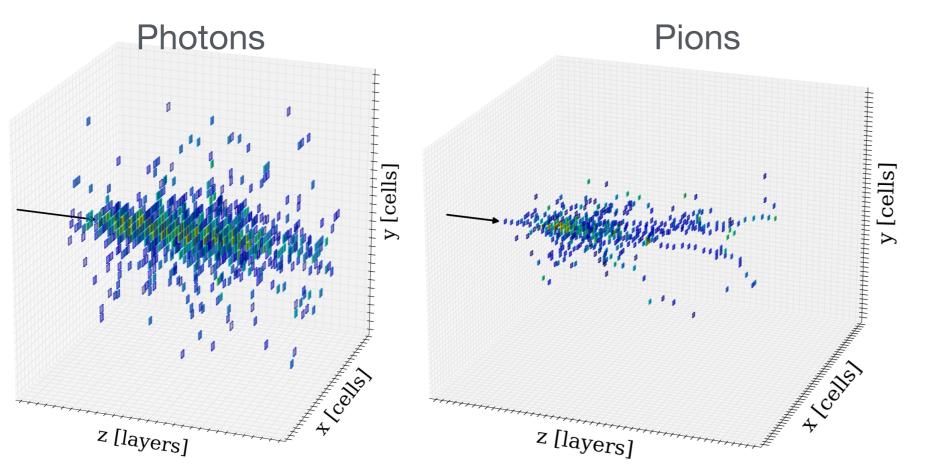
Machine Learning and LHC Event Generation

Anja Butter^{1,2}, Tilman Plehn¹, Steffen Schumann³, Simon Badger⁴, Sascha Caron^{5, 6} Kyle Cranmer^{7,8}, Francesco Armando Di Bello⁹, Etienne Dreyer¹⁰, Stefano Forte¹¹, Sanmay Ganguly¹², Dorival Gonçalves¹³, Eilam Gross¹⁰, Theo Heimel¹, Gudrun Heinrich¹⁴, Lukas Heinrich¹⁵, Alexander Held¹⁶, Stefan Höche¹⁷, Jessica N. Howard¹⁸, Philip Ilten¹⁹, Joshua Isaacson¹⁷, Timo Janßen³, Stephen Jones²⁰, Marumi Kado^{9,21}, Michael Kagan²², Gregor Kasieczka²³, Felix Kling²⁴, Sabine Kraml²⁵, Claudius Krause²⁶, Frank Krauss²⁰, Kevin Kröninger²⁷, Rahool Kumar Barman¹³, Michel Luchmann¹, Vitaly Magerya¹⁴, Daniel Maitre²⁰, Bogdan Malaescu², Fabio Maltoni^{28,29}, Till Martini³⁰, Olivier Mattelaer²⁸, Benjamin Nachman^{31,32}, Sebastian Pitz¹, Juan Rojo^{33,34}, Matthew Schwartz³⁵, David Shih²⁵, Frank Siegert³⁶, Roy Stegeman¹¹, Bob Stienen⁵, Jesse Thaler³⁷, Rob Verheyen³⁸, Daniel Whiteson¹⁸, Ramon Winterhalder²⁸, and Jure Zupan¹⁹



Butter, Plehn, Schumann, .. GK et al 2203.07460

Simulate showers initiated by photons and charged pions



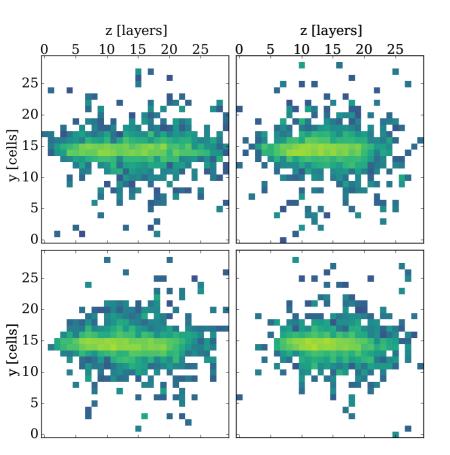
Hallin, .., GK et al 2109.00546;

Pions

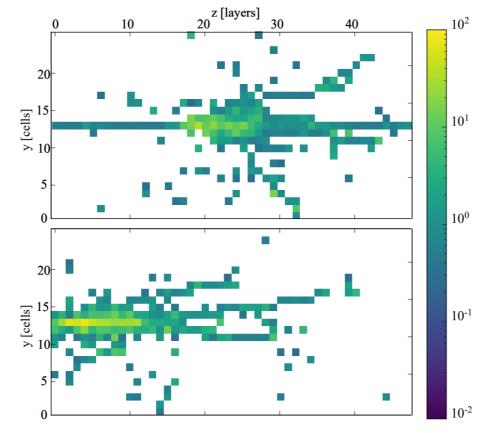
Simulate showers initiated by photons and charged pions

Pion showers are significantly more complex

Photons

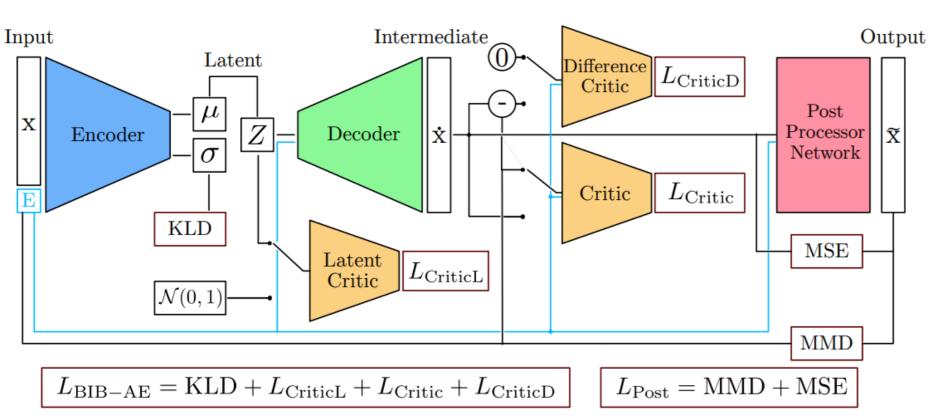


ILD ECal Silicon/Tungstate



ILD AHCal stainless steel absorber plates and scintillator tiles with SiPMs

Buhmann, ..., GK et al 2005.05334; Buhmann, ..., GK et al 2112.09709;



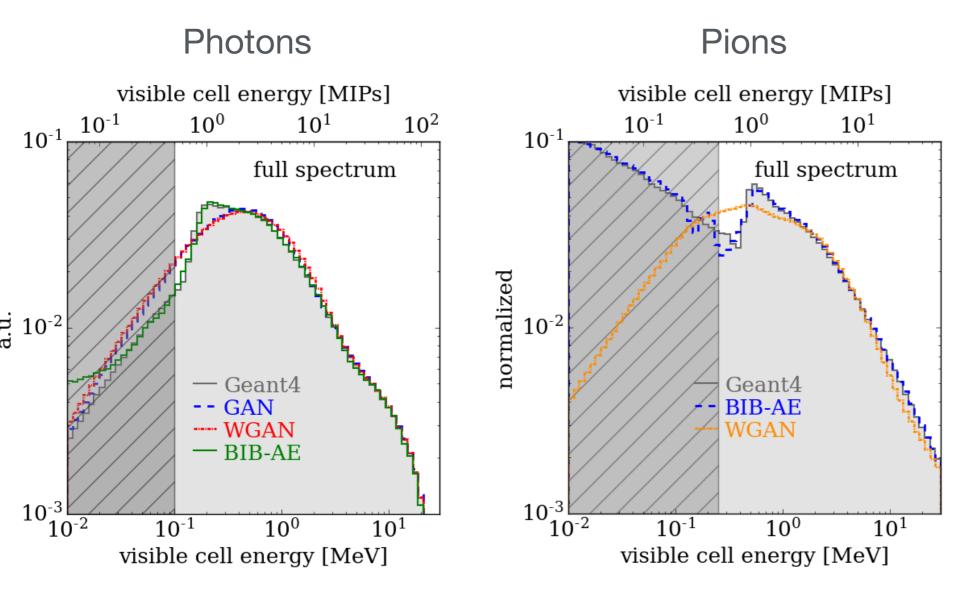
Simulate showers put initiated by photons and charged pions

> Pion showers are significantly more complex

Use modified BIB-AE architecture

$$\begin{split} L_{\text{BIB-AE}} &= -\beta_{C_L} \cdot \mathbb{E}_{x \sim p_{data(x)}} [C_L(E(x))] \\ &- \beta_C \cdot \mathbb{E}_{x \sim p_{data(x)}} [C(D(E(x)))] \\ &+ \beta_{\text{KLD}} \cdot \text{KLD}(E(x)) \\ &+ \beta_{\text{MMD}} \cdot \text{MMD}(E(x), \mathcal{N}(0, 1)). \end{split}$$

Buhmann, ..., GK et al 2005.05334; Buhmann, ..., GK et al 2112.09709; Voloshynovskiy et al 1912.00830;



Simulate showers initiated by photons and charged pions

Pion showers are significantly more complex

Use modified BIB-AE architecture

Achieves excellent modelling of differential distributions first for high-granularity calorimeter

Buhmann, .., GK et al 2005.05334; Buhmann, .., GK et al 2112.09709;

Timing

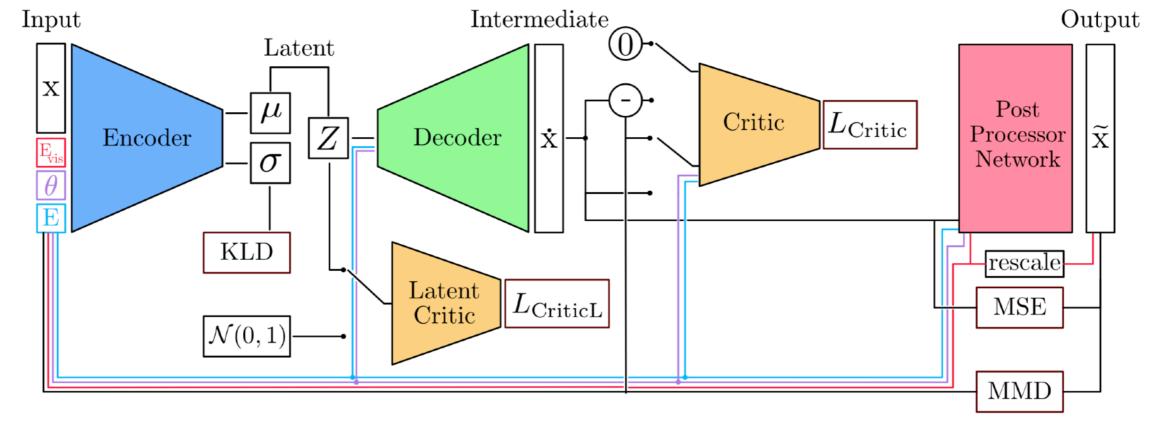
Simulator	Hardware	Batch Size	$15 { m GeV}$	Speed-up	10-100 GeV Flat	Speed-up
Geant4	CPU	N/A	$1445.05 \pm 19.34 \ {\rm ms}$	-	$4081.53 \pm 169.92 \ {\rm ms}$	-
WGAN	CPU	1	$64.34\pm0.58~\mathrm{ms}$	x23	$63.14\pm0.34~\mathrm{ms}$	x65
		10	$59.53\pm0.45~\mathrm{ms}$	$\mathbf{x24}$	$56.65 \pm 0.33 \text{ ms}$	x72
		100	$58.31\pm0.93~\mathrm{ms}$	$\mathbf{x25}$	$58.11 \pm 0.13 \text{ ms}$	x70
		1000	$57.99\pm0.97~\mathrm{ms}$	$\mathbf{x25}$	$57.99 \pm 0.18 \text{ ms}$	x70
BIB-AE	CPU	1	$426.60 \pm 3.27 \text{ ms}$	x3	$426.32 \pm 3.62 \text{ ms}$	x10
		10	$422.60\pm0.26~\mathrm{ms}$	x3	$424.71 \pm 3.53 \text{ ms}$	x10
		100	$419.64\pm0.07~\mathrm{ms}$	x3	$418.04 \pm 0.20 \text{ ms}$	x10
WGAN	GPU	1	$3.24\pm0.01~\rm{ms}$	x446	$3.25 \pm 0.01 \text{ ms}$	x1256
		10	$6.13\pm0.02~\mathrm{ms}$	x236	$6.13 \pm 0.02 \text{ ms}$	x666
		100	$5.43\pm0.01~\mathrm{ms}$	$\mathbf{x266}$	$5.43 \pm 0.01 \text{ ms}$	x752
		1000	$5.43\pm0.01~\mathrm{ms}$	x266	$5.43 \pm 0.01 \text{ ms}$	x752
BIB-AE	GPU	1	$3.14\pm0.01~\mathrm{ms}$	x460	$3.19\pm0.01~\mathrm{ms}$	x1279
		10	$1.56\pm0.01~\mathrm{ms}$	x926	$1.57\pm0.01~\mathrm{ms}$	x2600
		100	$1.42\pm0.01~\mathrm{ms}$	x1017	$1.42\pm0.01~\mathrm{ms}$	x2874

Photons

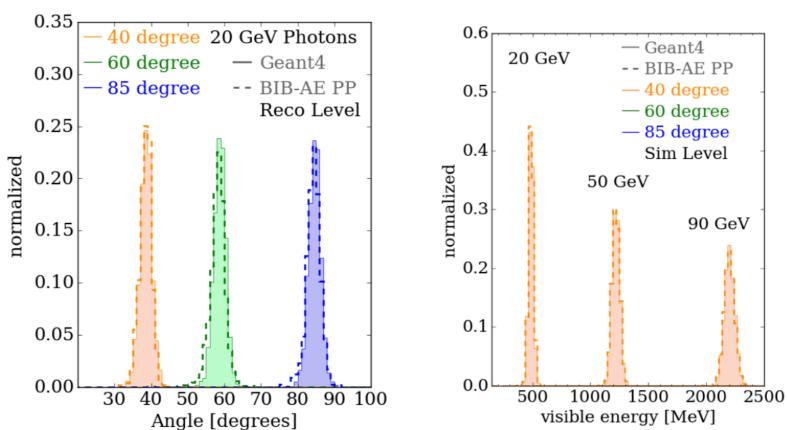
	Hardware	Simulator	Time / Shower [ms]	Speed-up
	CPU	Geant4	2684 ± 125	×1
Pions		WGAN BIB-AE	47.923 ± 0.089 350.824 ± 0.574	$\times 56 \\ \times 8$
	GPU	WGAN BIB-AE	$\begin{array}{c} 0.264 \pm 0.002 \\ 2.051 \pm 0.005 \end{array}$	$\begin{array}{c} \times 10167 \\ \times 1309 \end{array}$

Buhmann, .., GK et al 2005.05334; Buhmann, .., GK et al 2112.09709;

Dual conditioning



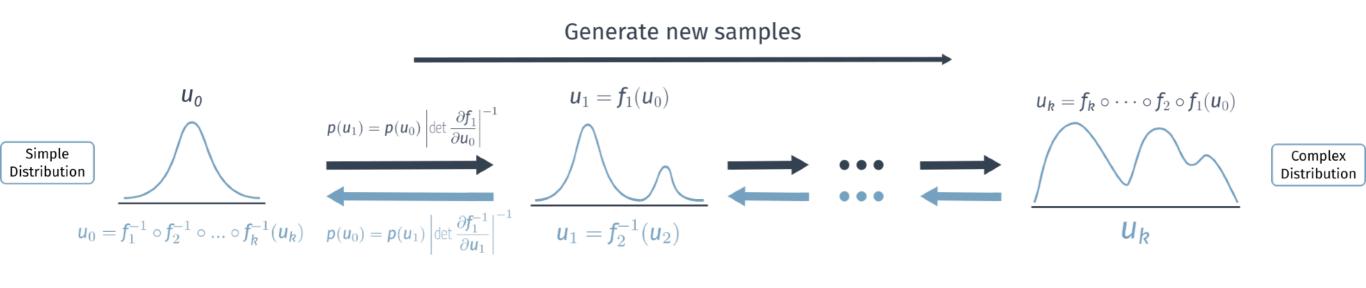
Additional control over one incoming particle angle



Diefenbacher .. GK et al 2303.18150

Flow based generation

How can we further improve the quality of generated distributions?



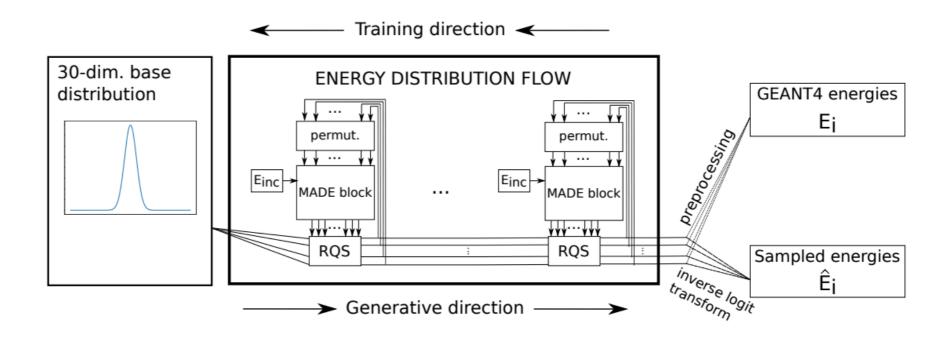
Evaluate probability/likelihood, train flow

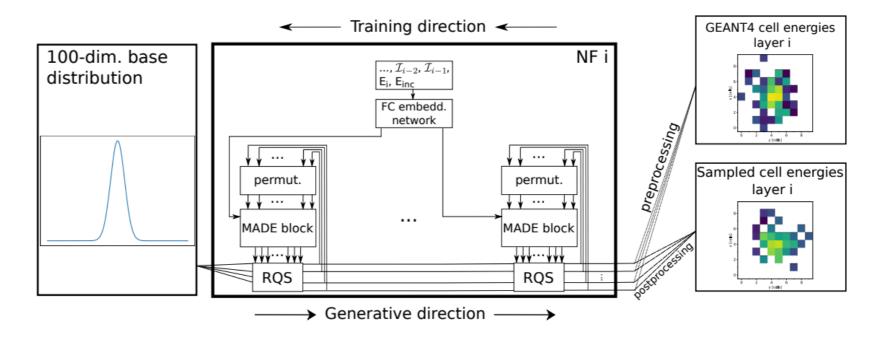
Flow based models learn an exact mapping to a Gaussian latent space, can be trained by direct likelihood minimisation (no adversarial terms, more stable!)

Flow based generation

Disadvantage: As invertible model, latent space dimension is same as data space dimension, bad scaling with large inputs

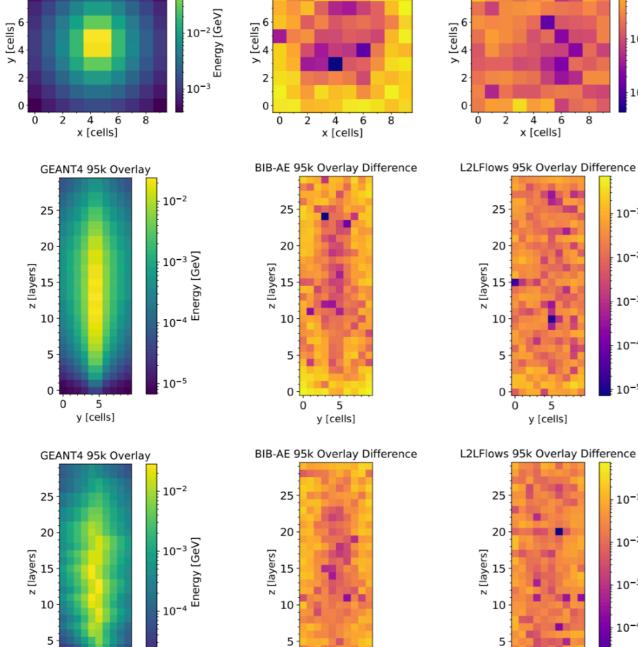
Solution: Two step generation, Sequential conditioning





Diefenbacher..GK et al 2302.11594; based on Shih/Krause CaloFlows I II

L2L Flows



0

Ó

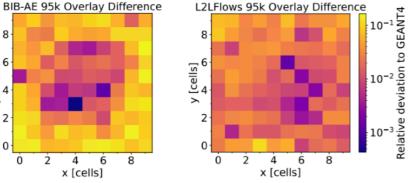
5

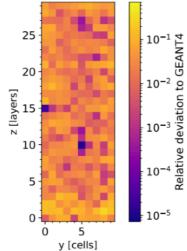
x [cells]

GEANT4 95k Overlay

8

10-1





GEANT4 10-1

deviation 10-3

Relative 10-4

10-2 5

10-5

25

20-

10 -

5

0 -

0

5

x [cells]

# Showers per simulator	AUC GEANT4 vs L2LFLOWS	AUC GEANT4 vs BIB-AE
95k	0.8518 ± 0.0042	0.9947 ± 0.0025
190k	0.8768 ± 0.0029	_
380k	0.8962 ± 0.0024	—
760k	0.9402 ± 0.0011	—

Better generative fidelity than BIB-AE, working on scaling up!

Diefenbacher..GK et al 2302.11594

10-5

0-

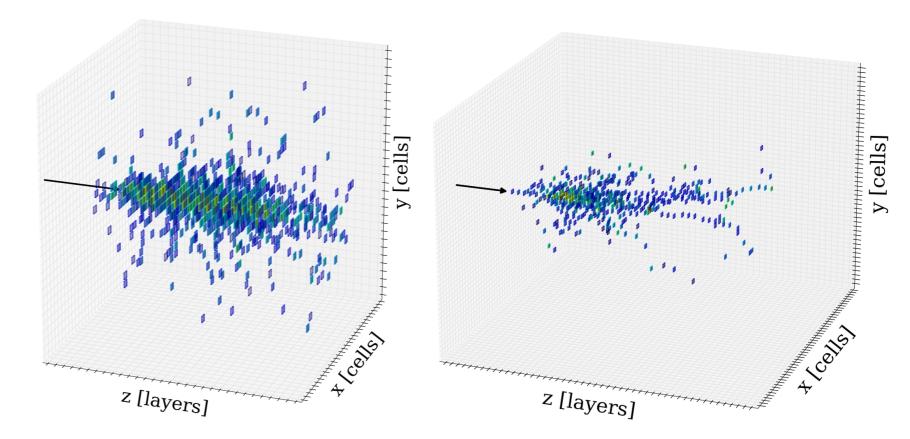
0

5

x [cells]

For high-resolution detectors, images will be very **sparse** and geometry **impossible** to map to fixed grid. Instead treat showers as **graphs or point clouds**

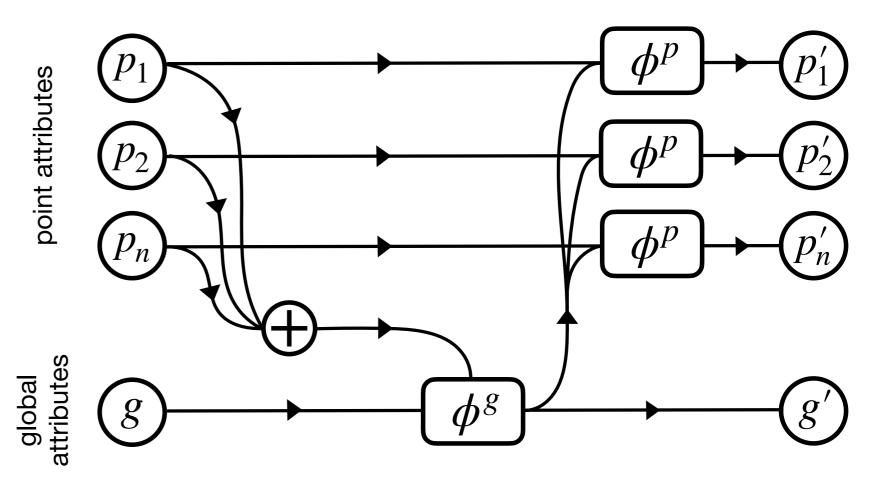
The physics of particle shower is **simple**: Point cloud (instead of graph) more efficient and **much faster**



Buhmann, GK, Thaler 2301.08128; Kansal et al 2106.11535; Käch et al 2211.13630; Buhmann, ... GK, et al 2305.04847

For high-resolution detectors, images will be very **sparse** and geometry **impossible** to map to fixed grid. Instead treat showers as **graphs or point clouds**

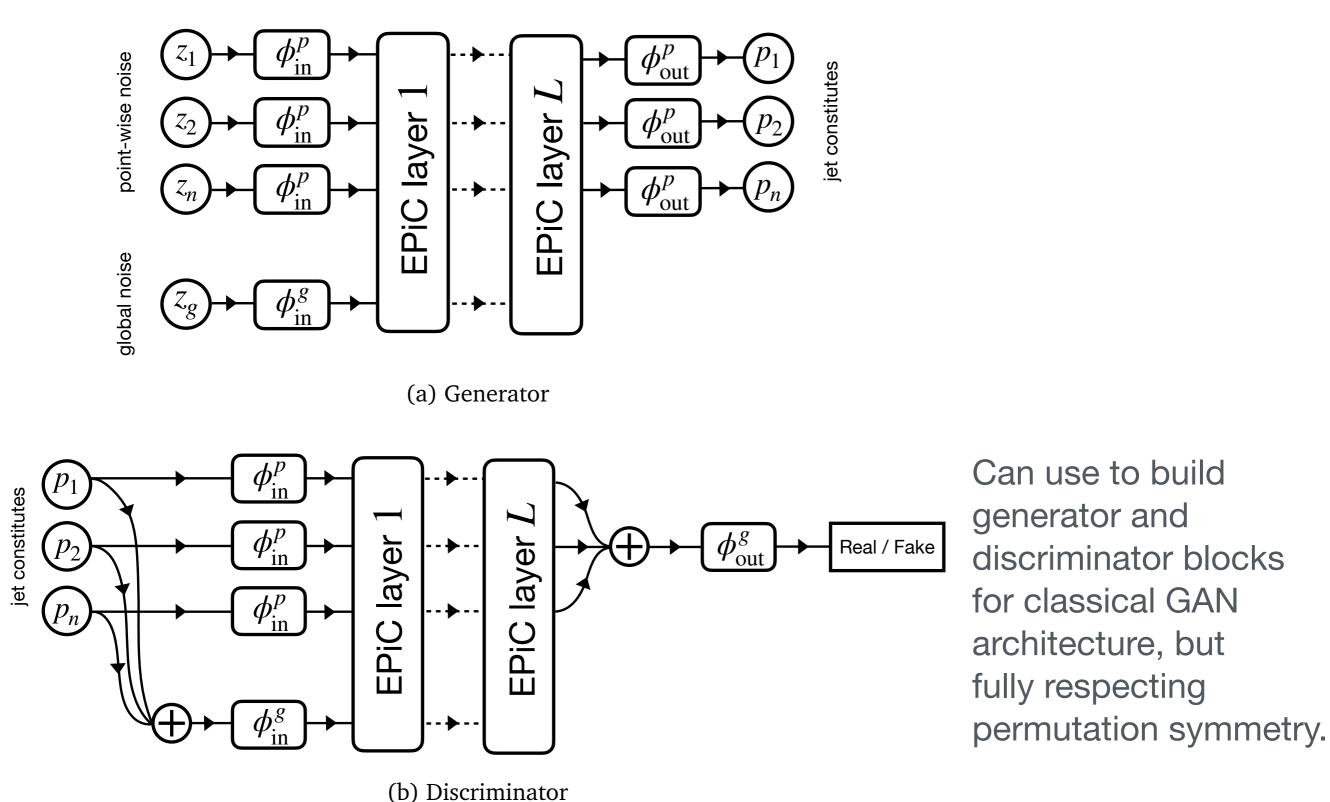
The physics of particle shower is **simple**: Point cloud (instead of graph) more efficient and **much faster**



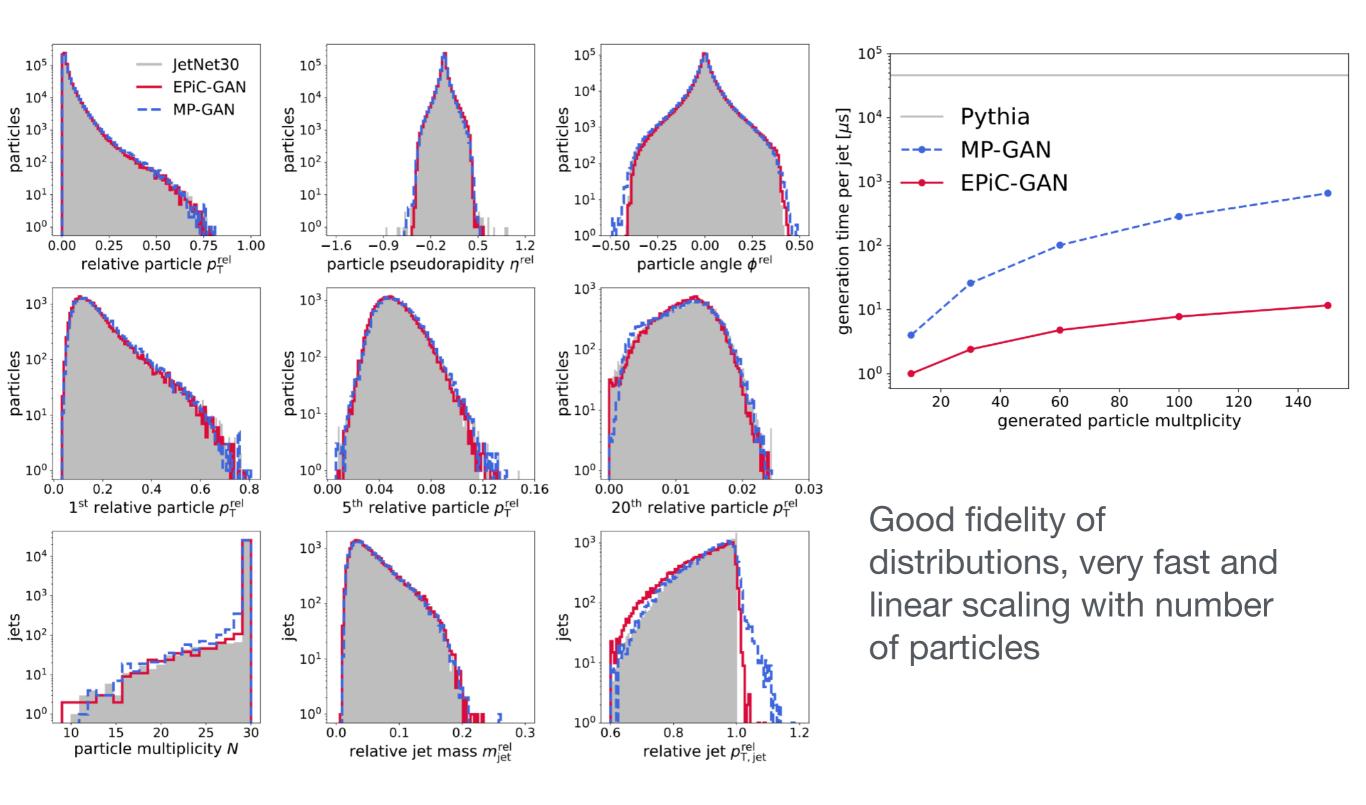
Equivariant Point Cloud interaction (EPiC) block: Similar to deep sets, but with additional global information exchange.

Still permutation equivariant

Buhmann, GK, Thaler 2301.08128; Kansal et al 2106.11535; Käch et al 2211.13630; Buhmann, ... GK, et al 2305.04847



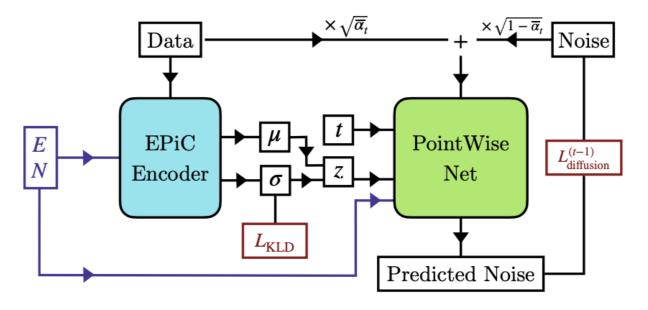
Buhmann, GK, Thaler 2301.08128;



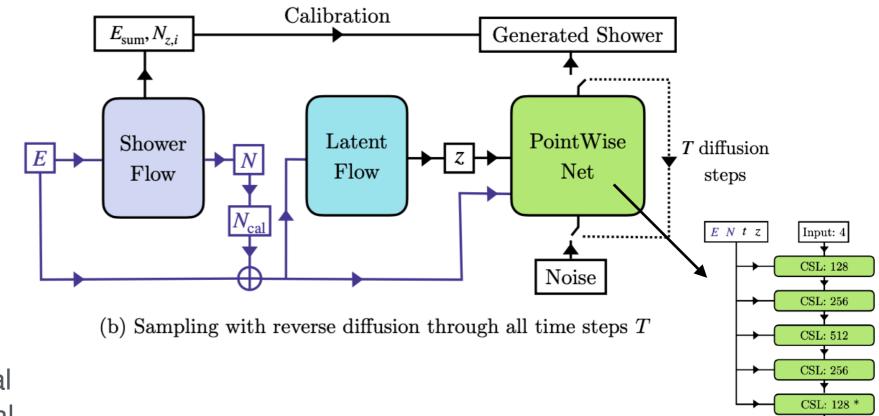
Buhmann, GK, Thaler 2301.08128;

To improve the generative fidelity, move from GAN to diffusion model:

Learn step-wise demonising



(a) Training at random time step t



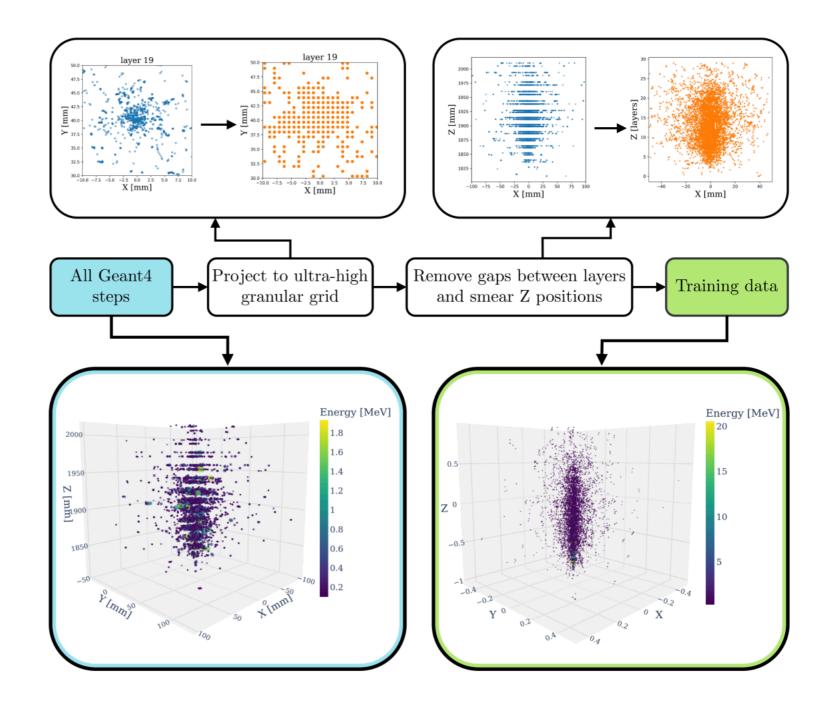
Output: 4

Buhmann, GK, Thaler 2301.08128; Kansal et al 2106.11535; Käch et al 2211.13630; Buhmann, ... GK, et al 2305.04847

To improve the generative fidelity, move from GAN to diffusion model:

Learn step-wise demonising

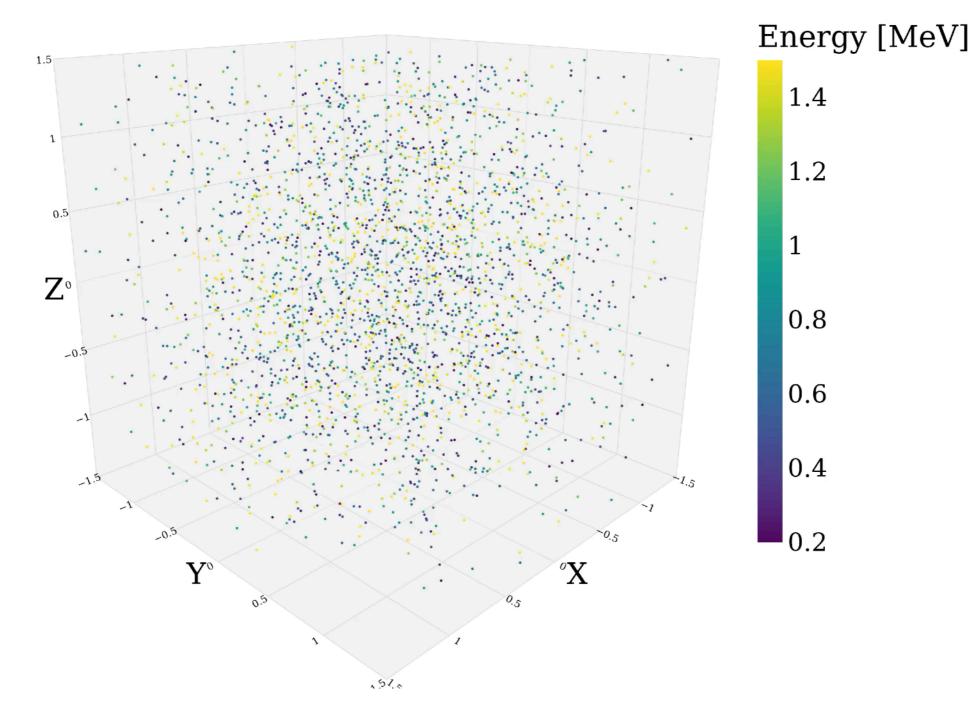
Some additional preprocessing needed



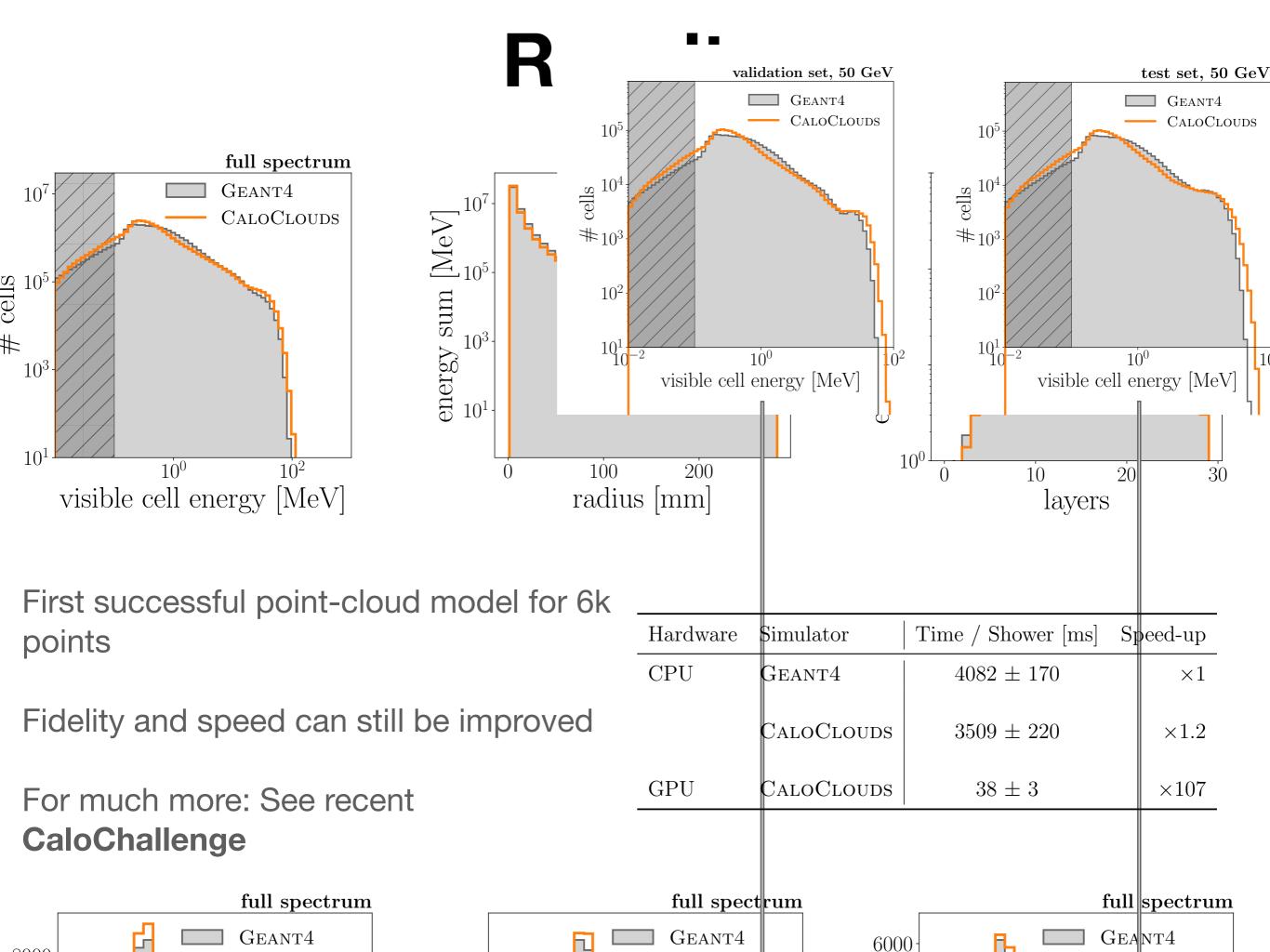
Buhmann, GK, Thaler 2301.08128; Kansal et al 2106.11535; Käch et al 2211.13630; Buhmann, ... GK, et al 2305.04847

Diffusion

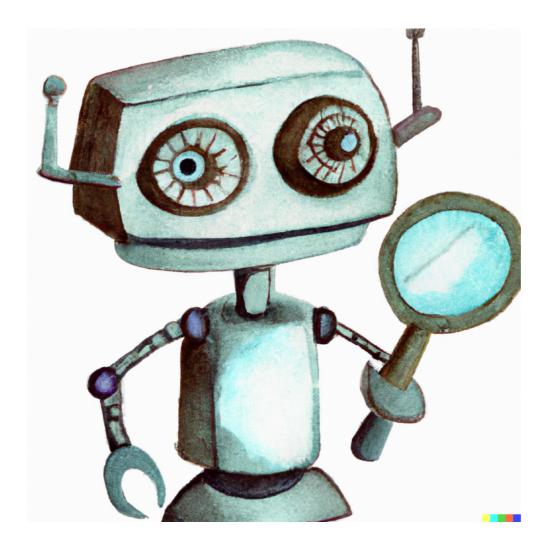
CaloCloud, time stamp: t_{99}



Buhmann, ... GK, et al 2305.04847

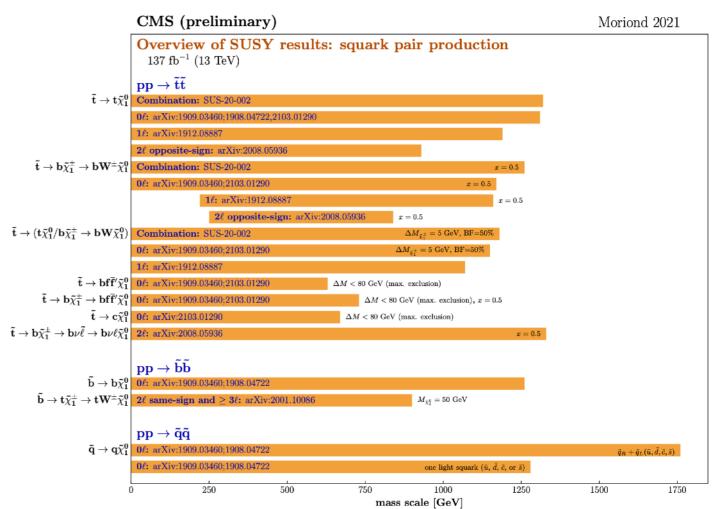


Discovering new Physics



Searches for new physics

- Theoretical and experimental reasons to expect new physics beyond the Standard Model
- However, so far only negative results in direct (model driven) searches
- Make sure that we do not miss potential discoveries at the LHC
 →Anomaly detection



Selection of observed limits at 95% C.L. (theory uncertainties are not included). Probe up to the quoted mass limit for light LSPs unless stated otherwise. The quantities ΔM and x represent the absolute mass difference between the primary sparticle and the LSP, and the difference between the intermediate sparticle and the LSP relative to ΔM , respectively, unless indicated otherwise.

What is an anomaly?



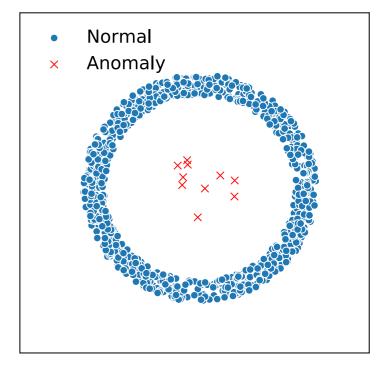
Point anomaly

- Outliers: Datapoints far away from regular distribution
- Examples:
 - Detector malfunctions
 - Background-free search









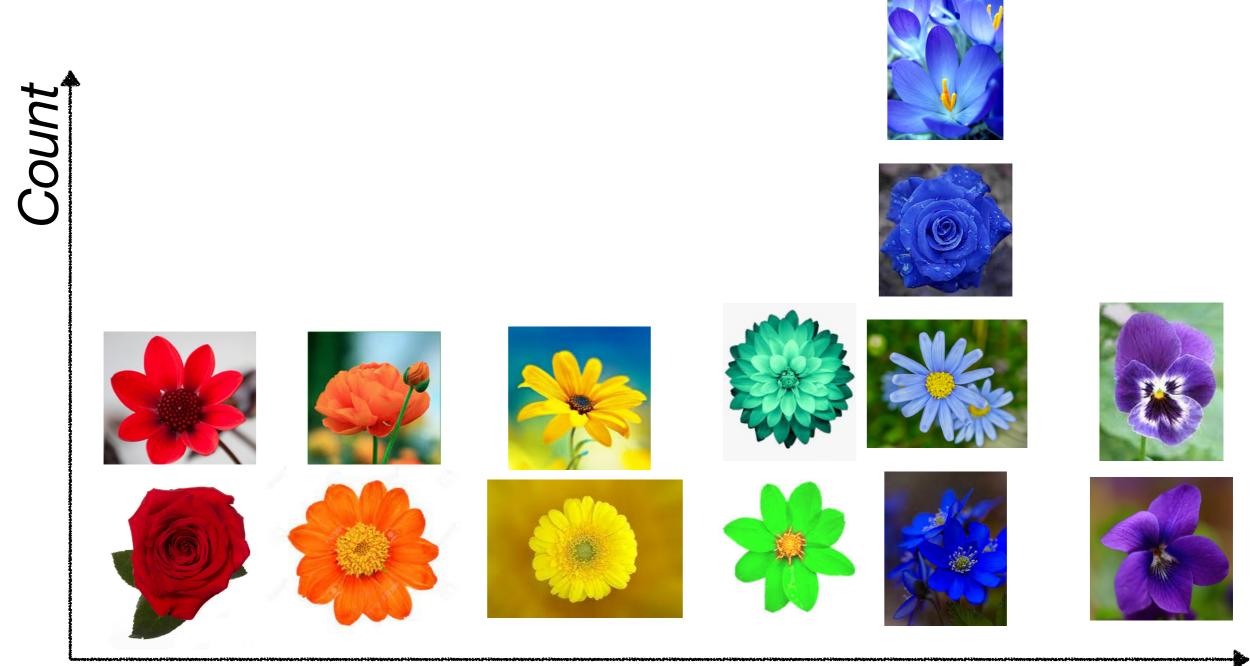




And now?



Group anomaly

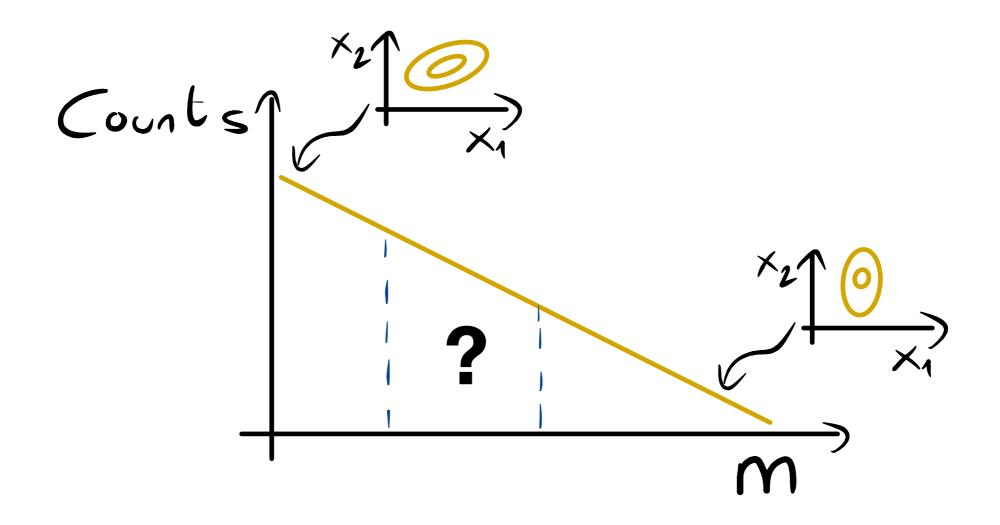


Color (Mass)

Group anomaly

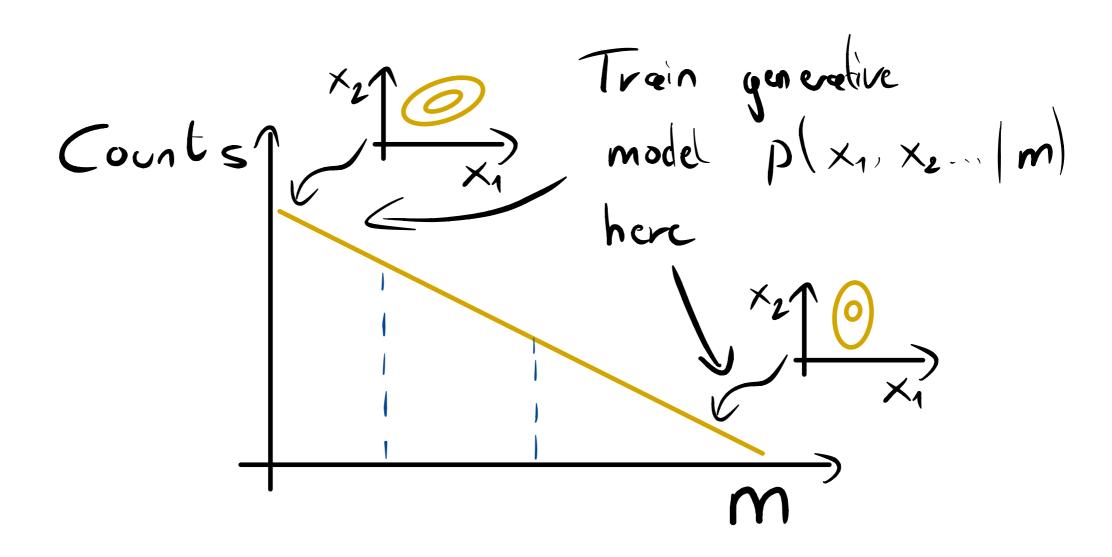
- Individual examples not anomalous, Count but interesting collective behaviour a.u. Examples: New physics searches, e.g. resonances Excess in time series SB SR SB m $p_{\text{data}}(x|m \in SB)$ $p_{\text{data}}(x|m \in SB)$ $p_{\text{data}}(x|m \in SR)$ $= p_{bg}(x|m \in SB)$ $= p_{bg}(x|m \in SB)$
 - Color (Mass)

Shape Interpolation



Can learn distributions in arbitrary dimensions (shape) as long as interpolation in one variable

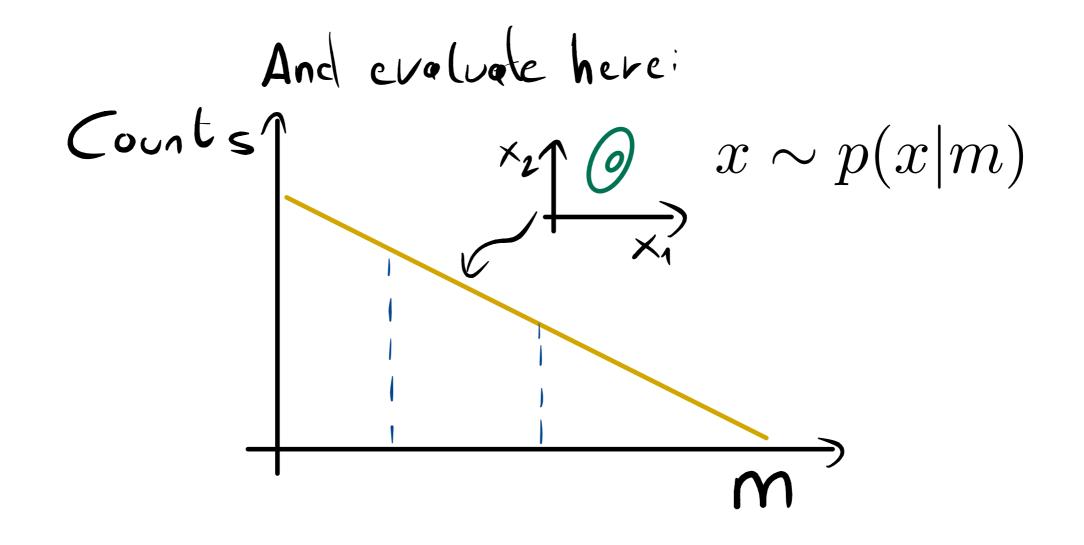
Shape Interpolation



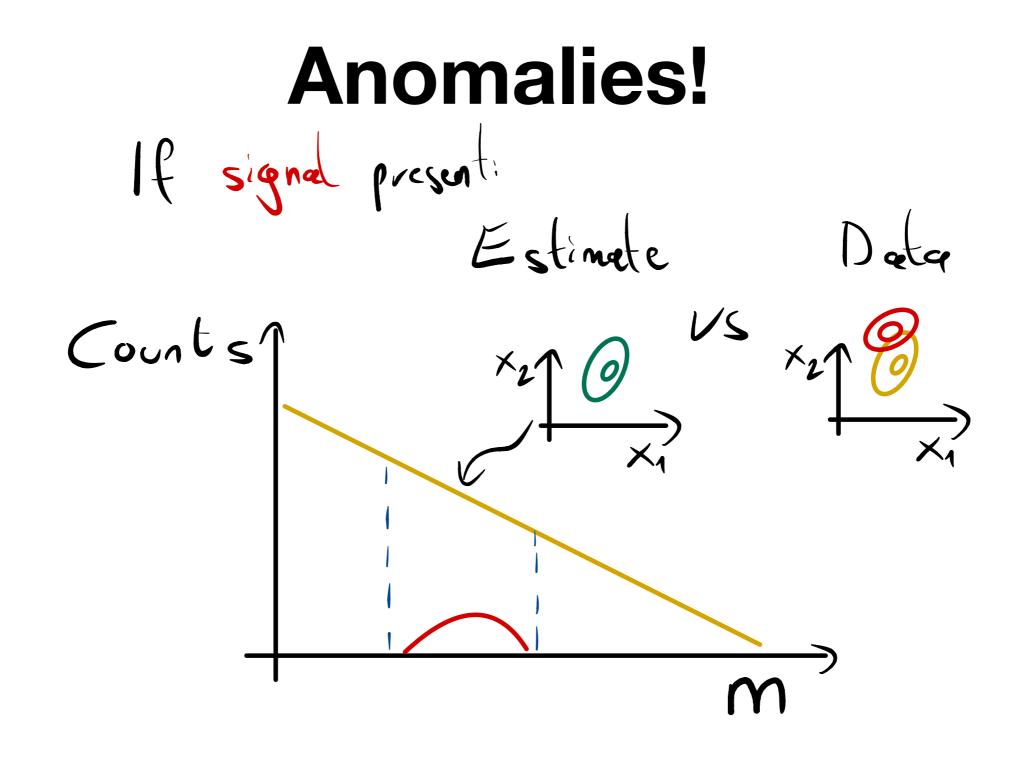
Can be any conditional generative model (GAN, VAE, flow, diffusion)

In practice we use a normalising flow

Shape Interpolation



Sampling yields an **in-situ background** estimate trained on data



Training a binary classifier between estimate and data can detect localised signals (e.g. resonant new physics)

Aside: Justification

Per Neyman-Pearson: Likelihood-ratio is optimal test statistic *Unfortunatly, p(x|anomaly) is not available*

$$L_{S/B} = \frac{p(x|\text{anomaly})}{p(x|\text{normal})}$$

$$L_{D/B} = \frac{p(x)}{p(x|\text{normal})}$$

Build data/background ratio:

Approximate background density using control measurement (e.g. sideband)

$$L_{D/B} \approx \frac{p(x)}{\tilde{p}(x|\text{normal})}$$

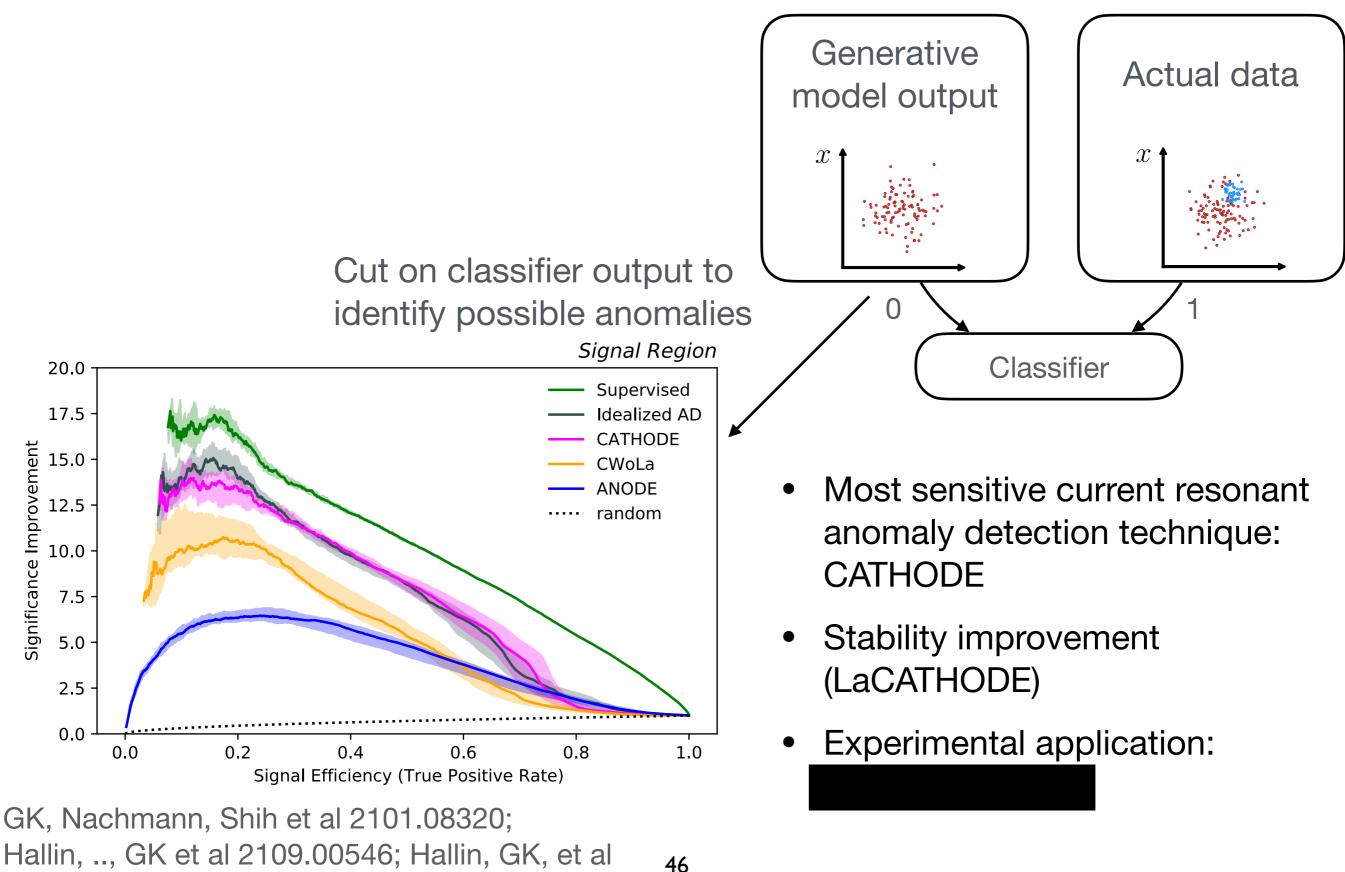
Expand

$$p(x) = f_{\text{normal}} p(x|\text{normal}) + f_{\text{anomaly}} p(x|\text{anomaly})$$

And insert:

$$L_{D/B} \approx f_{\text{normal}} + f_{\text{anomaly}} \frac{p(x|\text{anomaly})}{\tilde{p}(x|\text{normal})}$$

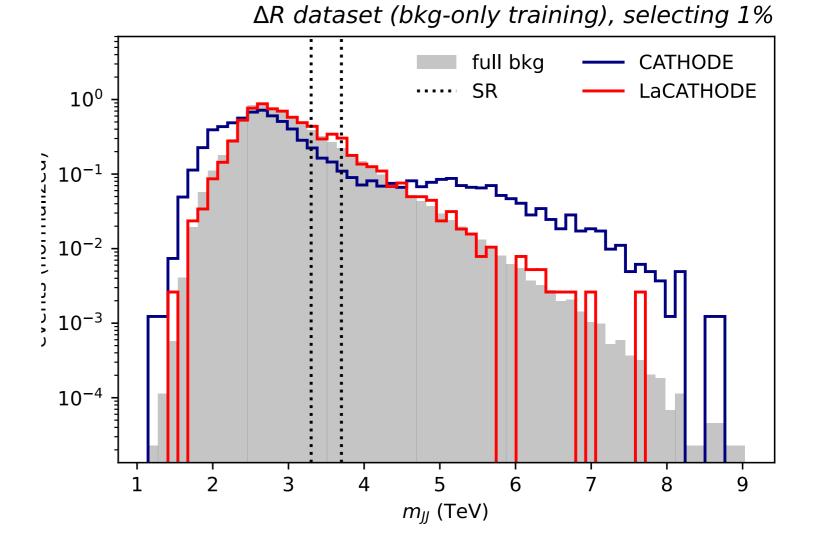
CATHODE & Friends



2210.14924

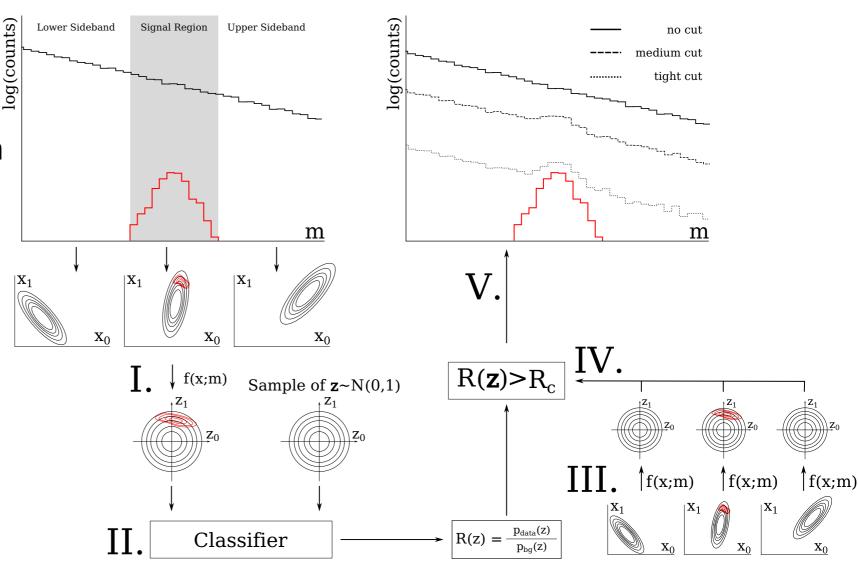
LaCATHODE

- If R(x) is only calculated in signal region, it's extrapolation is not well-defined
- Potential problem for bumphunt if it shapes distributions



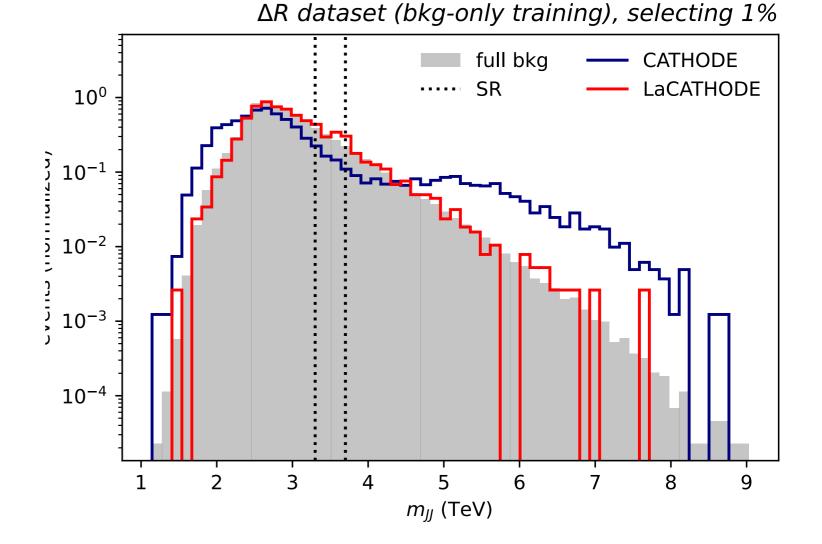
LaCATHODE

- If R(x) is only calculated in ^B
 signal region, it's extrapolation is not well-defined
- Potential problem for bumphunt if it shapes distributions
- Instead, train classifier in latent space



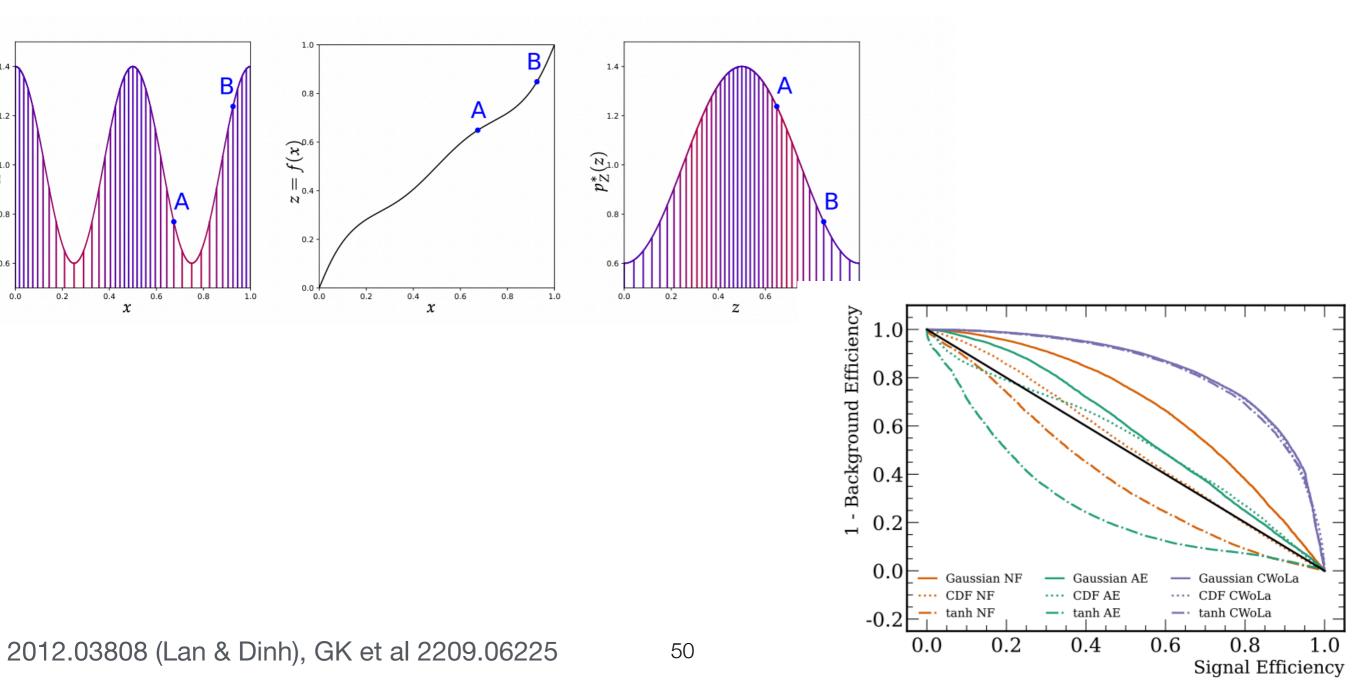
LaCATHODE

- If R(x) is only calculated in signal region, it's extrapolation is not well-defined
- Potential problem for bumphunt if it shapes distributions
- Instead, train classifier in latent space to achieve flat distributions



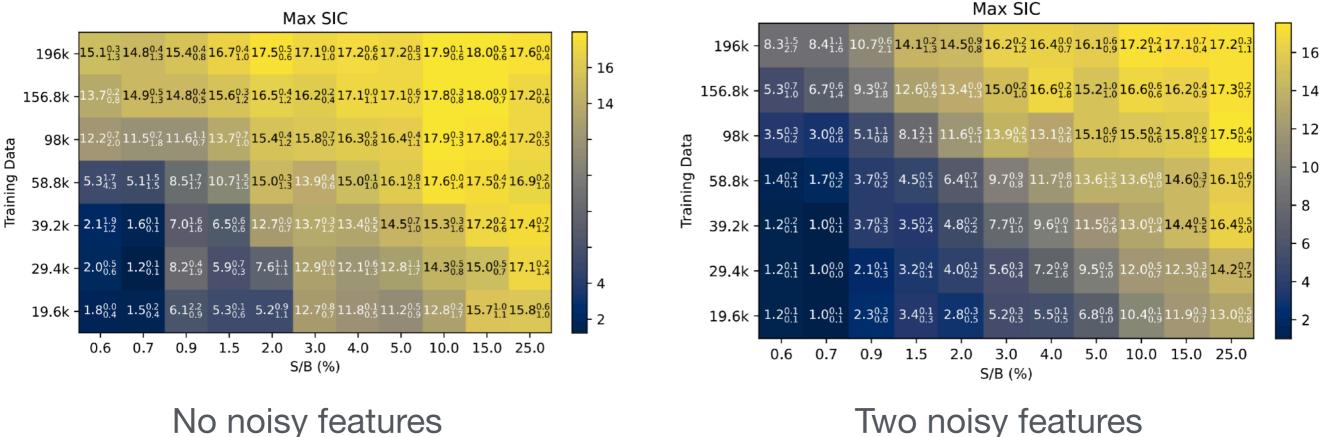
Comments on anomaly detection

• As CATHODE approximates a likelihood ratio, it should be robust compared to methods that only use p_{Background} (e.g. autoencoders)



Comments on anomaly detection

- As CATHODE approximates a likelihood ratio, it should be robust compared to methods that only use pBackground (e.g. autoencoders)
- However, still can be sensitive to choice of input features
 - Here shown: idealised anomaly detector (perfect DE)



51

Two noisy features

Closing



Advertisment

ML4Jets2023

Consider coming to Hamburg!

ML4Jets from November 6-10

Registration and Abstract submission are **open** under

https://indico.cern.ch/e/ml4jets

6–10 Nov 2023 DESY Europe/Zurich timezone		Enter your search term	٩	
Overview	Machine learning has become a hot topic in particle physics over the	e past several years. In particula	аr,	
Call for Abstracts	there has been a lot of progress in the areas of particle and event identification, reconstruction, generative models, anomaly detection and more. In this conference, we will discuss current progress in these areas, focusing on new breakthrough ideas and existing challenges. The ML4Jets workshop will			
Timetable	be open to the full community and will include LHC experiments as phenomenologists interested in this topic. Contributions from metho	ll as theorists and		
Registration	scientific fields facing similar challenges are welcome as well.			
Participant List	This year's conference is organised jointly by DESY and Universität F campus. It follows conferences in 2017, 2018, 2020, 2021, and 2022	-	Y	
Code of Conduct	•			
Contact Ml4jets2023-info@desy.de	Registration and abstract submission are open now and close on October 6th (registration for in- person participation) and September 10 (abstracts).			
	The workshop will be organised in a hybrid format (with a Zoom connection option). We expect speakers to attend in-person.			
	Registration for both in-person and Zoom-participation will be free of charge and (at the minimum) include coffee-breaks for in-person participants. We are looking into an opt-in dinner and announce details and potential extra costs closer to the event.			
	Local Organizing Committee: Freya Blekman (DESY & Universität Hamburg) Andrea Bremer (Universität Hamburg) Frank Gaede (DESY)			
	Gregor Kasieczka (Universität Hamburg, chair) Andreas Hinzmann (DESY)			
	Matthias Schröder (Universität Hamburg)			
	International Advisory Committee: Florencia Canelli (University of Zurich) Kyle Cranmer (NYU) Vava Gligorov (LPNHE) Gian Michele Innocenti (CERN) Ben Nachman (LBNL) Mihoko Nojiri (KEK) Maurizio Pierini (CERN) Tilman Plehn (Heidelberg) David Shih (Rutgers) Jesse Thaler (MIT)			
	Sofia Vallescorsa (CERN)			

Closing

- Deep Learning for particle physics is rapidly developing solutions to a wide range of problems
 - Classification
 - Anomaly detection
 - **Robustness and uncertainties** lacksquare
 - Efficient generation
 - Fast processing
- Physics encounters challenges of complex data and large volumes with potential relevance to other domains

Contact:

- Email: gregor.kasieczka@uni-hamburg.de
- Twitter: @GregorKasieczka
- Webpage: <u>https://www.physik.uni-hamburg.de/en/iexp/gruppe-</u> kasieczka.html Thank you

