Application of Machine-Learning Based Generative Modelling in Real-World Systems

Denis Derkach

LAMBDA laboratory, HSE university

LAMBDA • HSE

September 11, 2024

Generative Modeling

This X Does Not Exist!

This Person Does Not Exist

The site that started it all, with the name that says it all. Created using a style-based generative adversarial network (StyleGAN), this website had the tech community buzzing with excitement and intrigue and inspired many more sites.

Created by Phillip Wang.

This Cat Does Not Exist

These purr-fect GAN-made cats will freshen your feeline-gs and make you wish you could reach through your screen and cuddle them. Once in a while the cats have visual deformities due to imperfections in the model – beware, they can cause nightmares.

Created by Ryan Hoover.

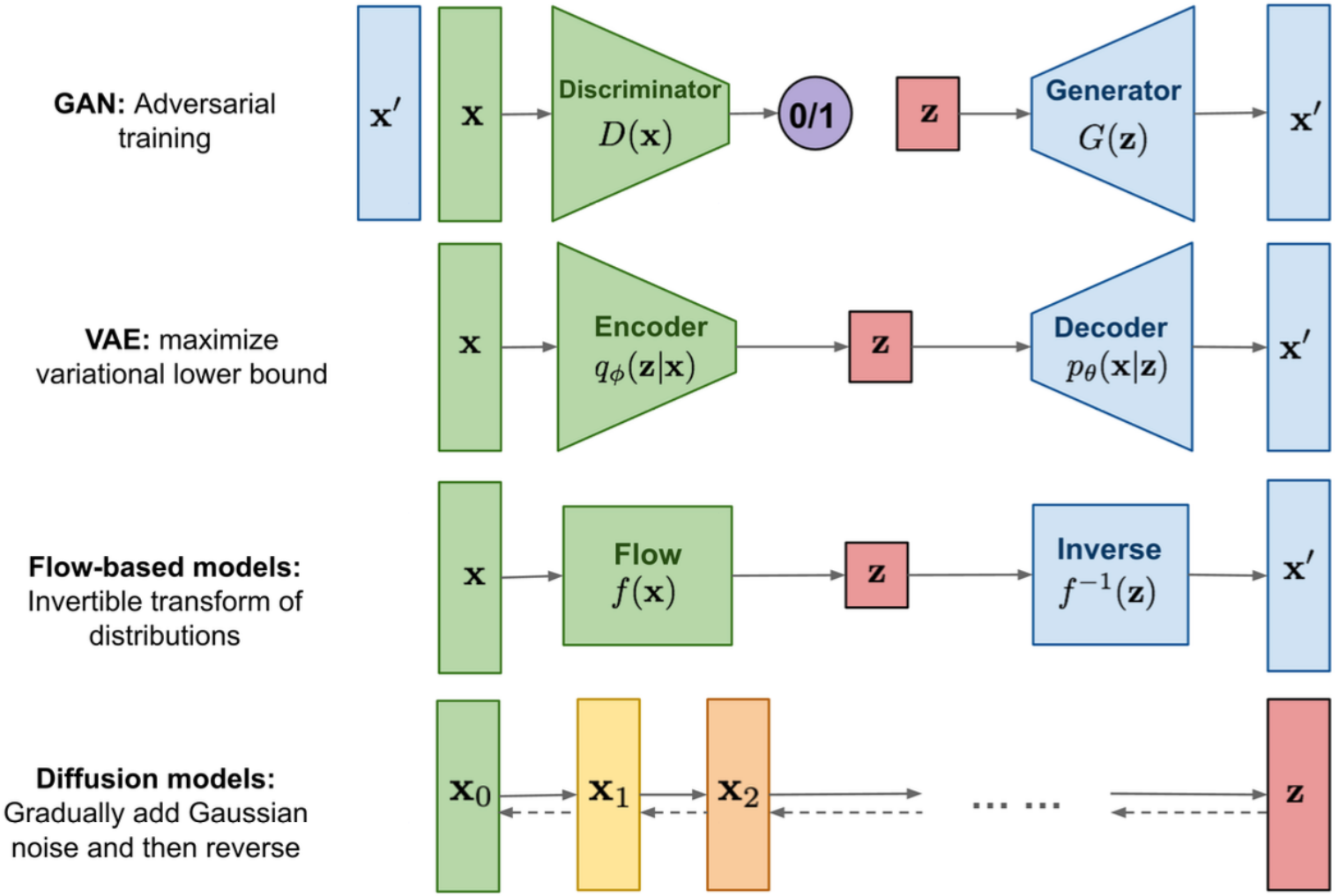
This Rental Does Not Exist

Why bother trying to look for the perfect home when you can create one instead? Just find a listing you like, buy some land, build it, and then enjoy the rest of your life.

Created by Christopher Schmidt.

https://thisxdoesnotexist.com/

Generative model: Model Zoo

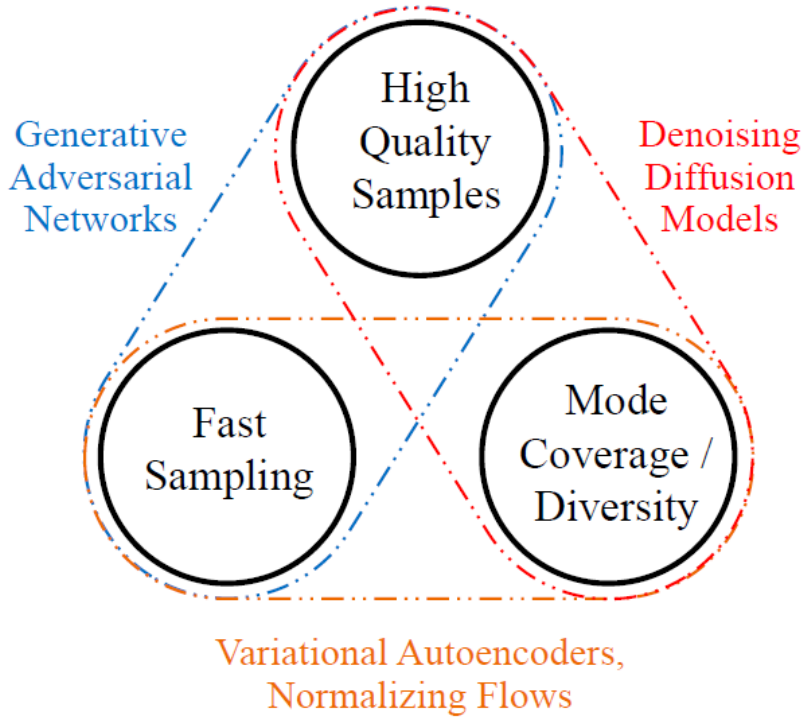


https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

Generative model: Problem Statement

Three major tasks, given a generative model f from a class of models $\mathcal{F}$:

- **Estimation**: find the f in $\mathcal{F}$ that best matches observed data.
- **Evaluate Likelihood**: compute $f(z)$ for a given z.
- **Sampling**: drawing from f.



S. Nowozin et al. f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization

Z. Xiao et al., Tackling the Generative Learning Trilemma with Denoising Diffusion GANs

Generative Models Tricks for Brains

- Enormous progress in recent years.
- Mostly tricks for brains.
- Technology is ready for new tasks.

https://davidleonfdez.github.io/gan/2022/05/17/gan-convergence-stability.html

Generative Models Tricks for Brains

- Enormous progress in recent years
- Mostly tricks for brains.
- Technology is ready for new tasks.

https://davidleonfdez.github.io/gan/2022/05/17/gan-convergence-stability.html

FAST COMPANY

02-08-19

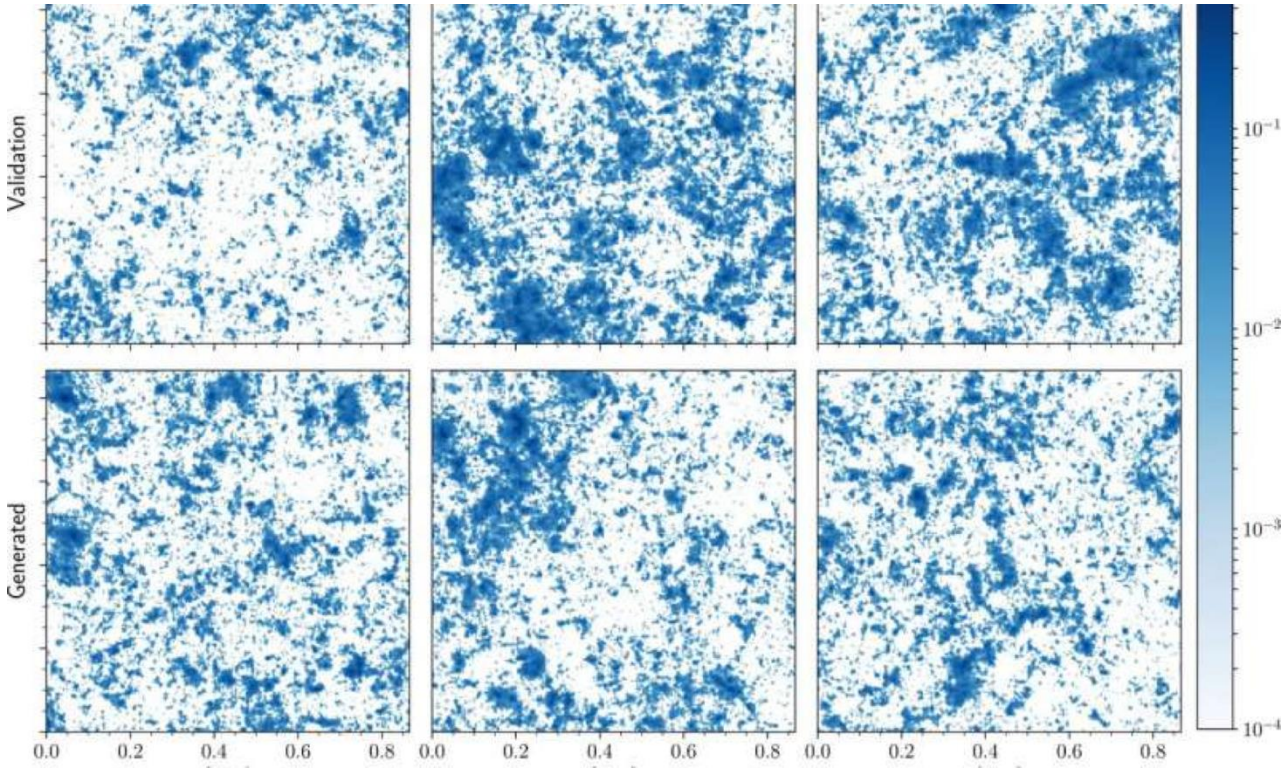
This AI dreams about cats–and they'll haunt your nightmares

Nvidia's new AI is capable of generating everything from human faces to kittens. But the development process left behind plenty of. . .errors.

https://www.fastcompany.com/90303908/this-ai-dreams-about-cats-and-theyll-haunt-your-nightmares

Astronomy Example

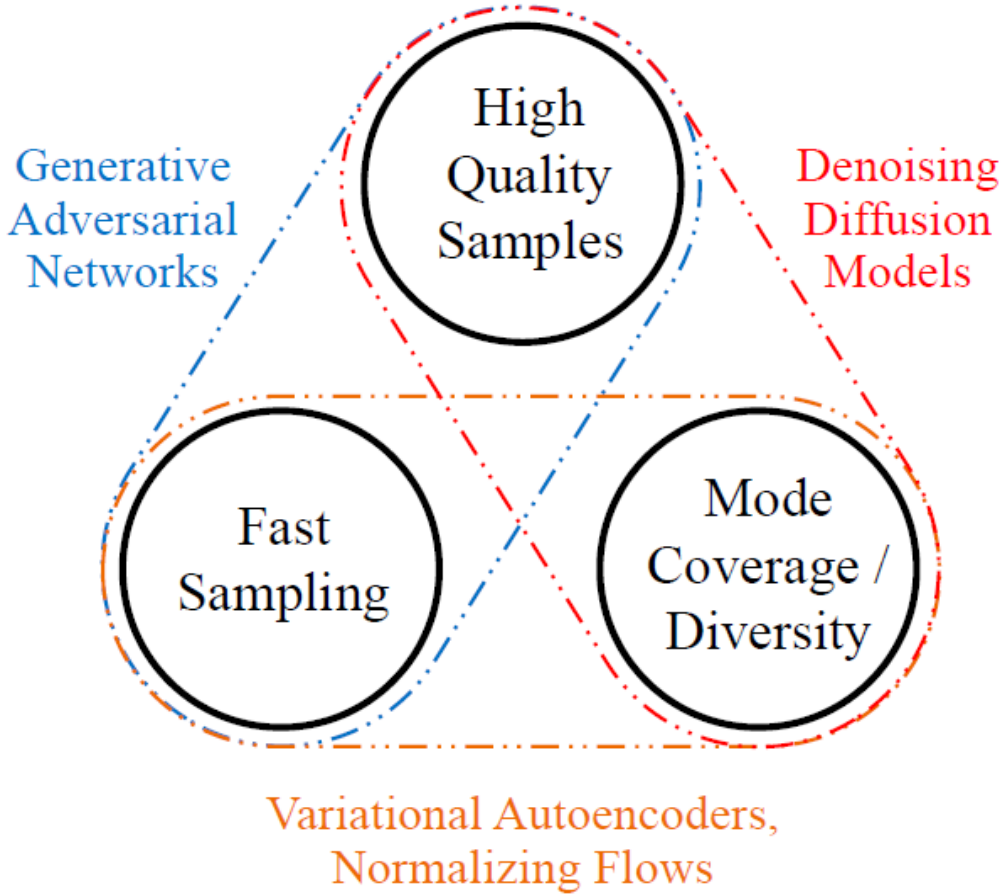
- Generate weak lensing convergence maps.
- "Visually, an **expert** cannot distinguish the generated maps from the full simulation ones"



Mustafa, M., et al.. Comput. Astrophys. 6, 1 (2019).

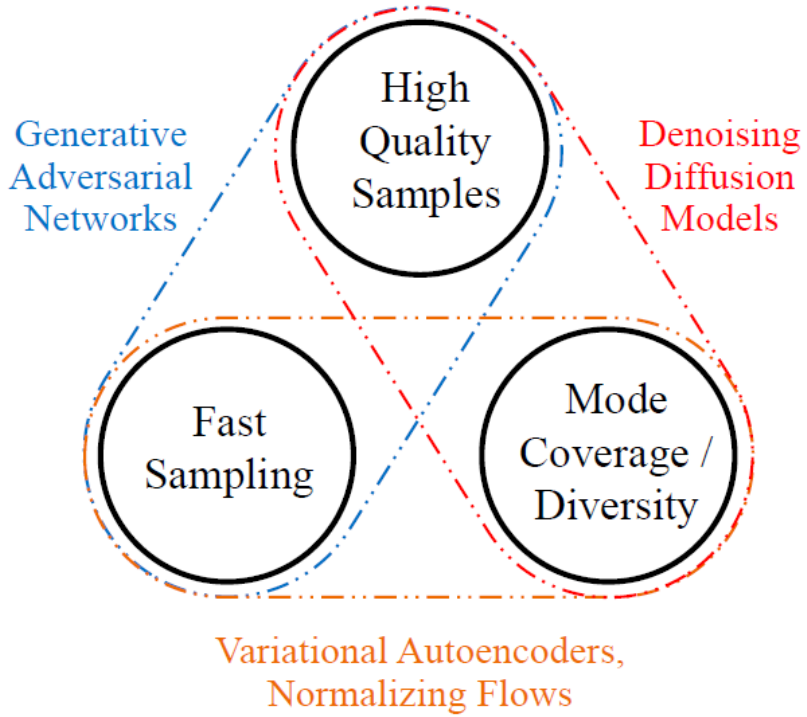
Generative models: take home message

- No single best generative model.
- Choices have consequences.
- The choice is problem motivated: cannot get all three at once.



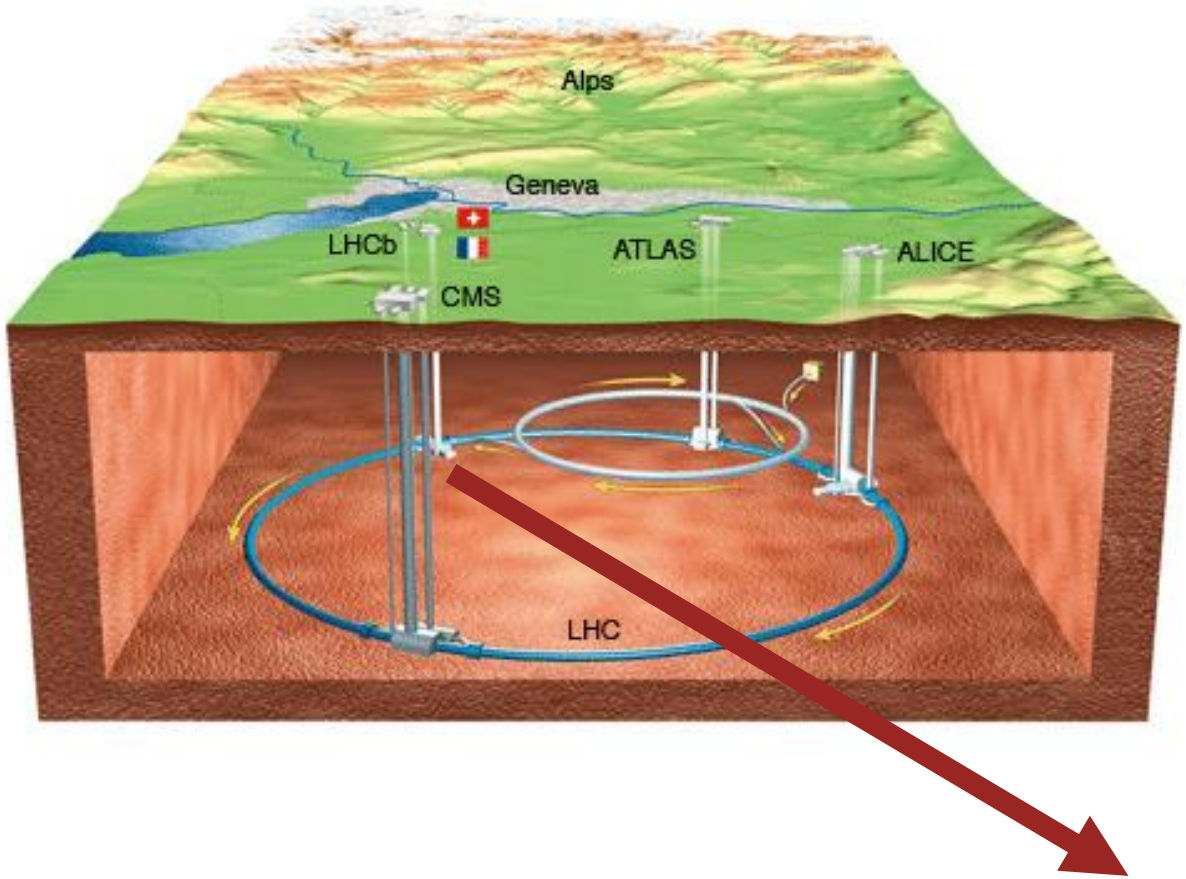
Fast Sampling

Fast Sampling Statement

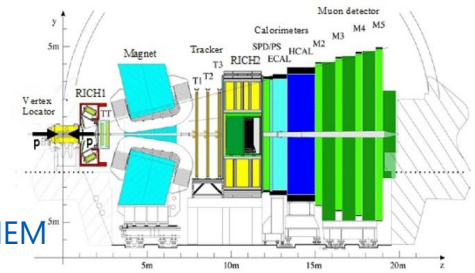


- Need a fast and flexible approach to generate as many realistic samples as possible.
- Trained on real data.
- With a possibility to correct the model with new data.

Large Hadron Collider



- Need “cameras” (or detectors) to see what happened.
- Millions of events (up to 4Tb/s data stream).
- Need to simulate known events.



High-Energy Physics

- Event can be considered as a photo.
- The event is than passed through the pipeline.

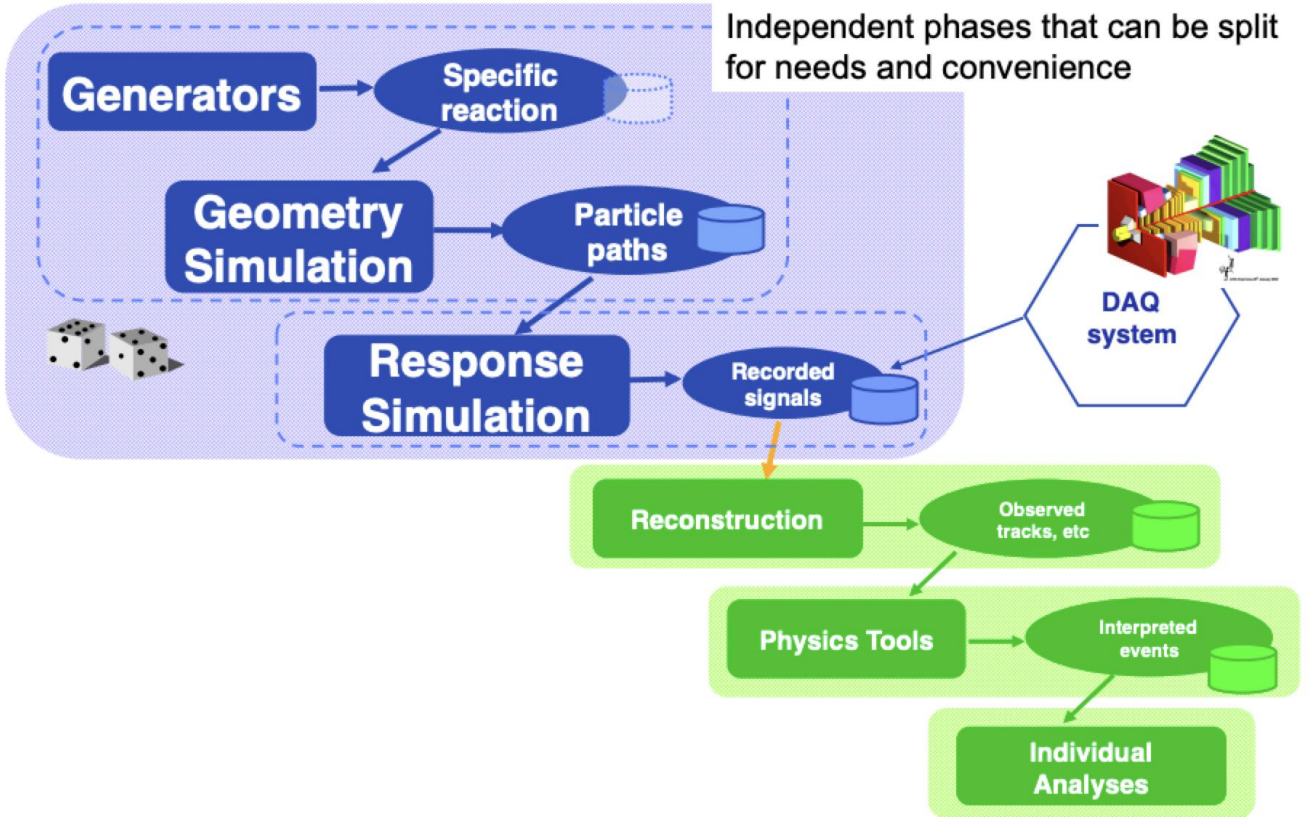
Simulation

Several model-motivated transitions.

Sequence:

- collision;
- decay;
- matter interaction;
- digitisation;
- reconstruction.

Each event takes 1 minute to generate (real world data is “generated” at several MHz).



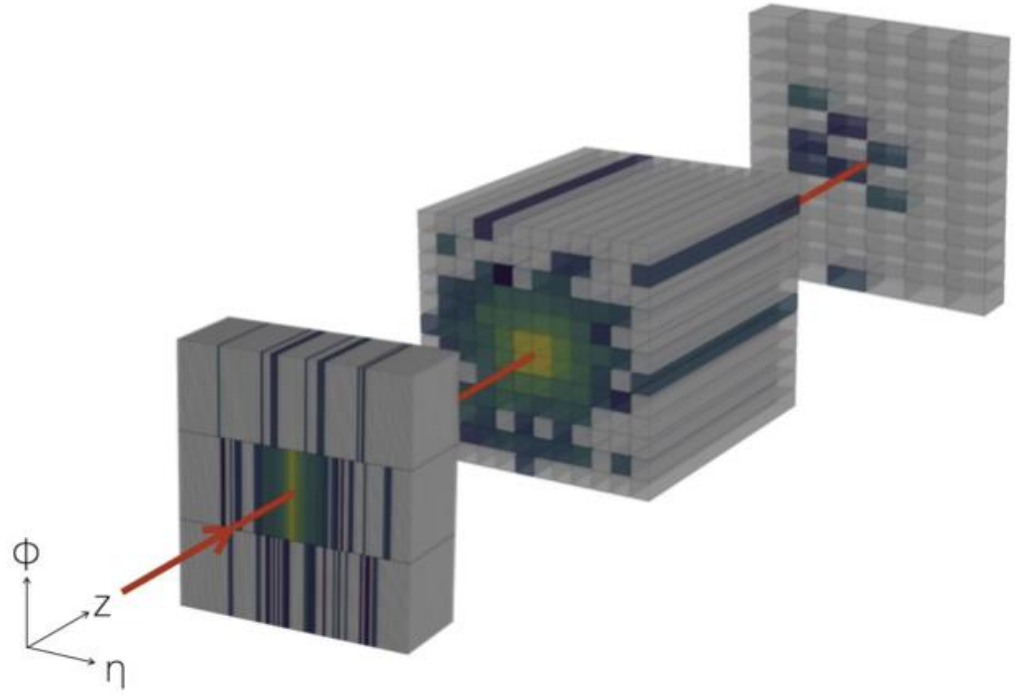
Simulate “simulation” using effective parameterization.

What is the event.

The calorimeter consists of many cells that reads out the energy deposit of a single particle.

A single particle deposits energy to several cells. An event is a sum of all particles and some noise.

We are normally using some reconstructed parameters of the event.



Paganini, M. et al. "CaloGAN: Simulating 3D high energy particle showers in multilayer electromagnetic calorimeters with generative adversarial networks." *Physical Review D* 97.1 (2018): 014021.

Ideas for Simulation

Since we know all processes in the subdetector, we can fully simulate an event using precise physics-motivated rules.

For calorimeters this means taking into account the structure of response that consists of many secondary particles.

This is done using Geant toolkit.

Pro: physics behind the simulation is controlled
Cons: slow, needs fine tuning.

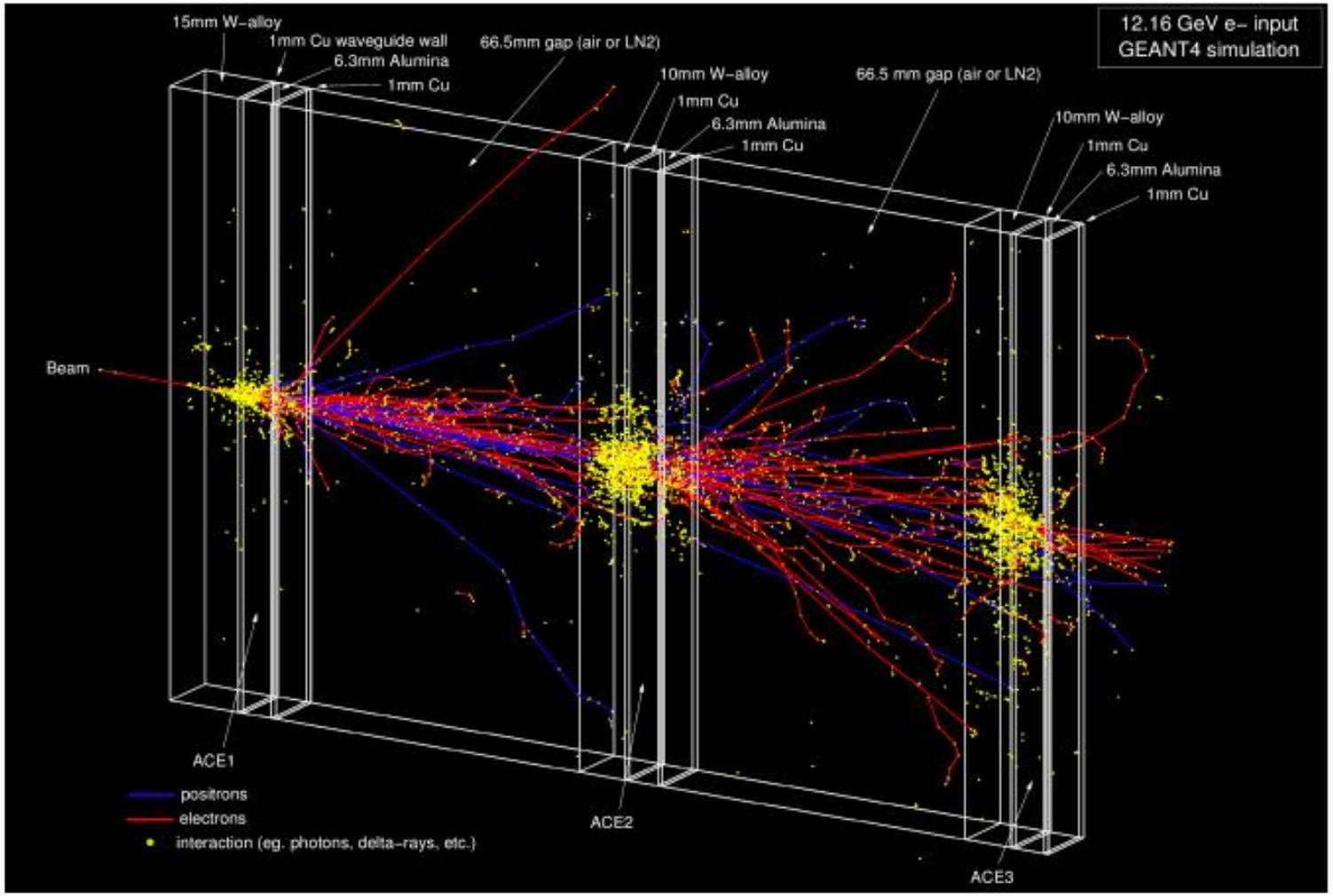
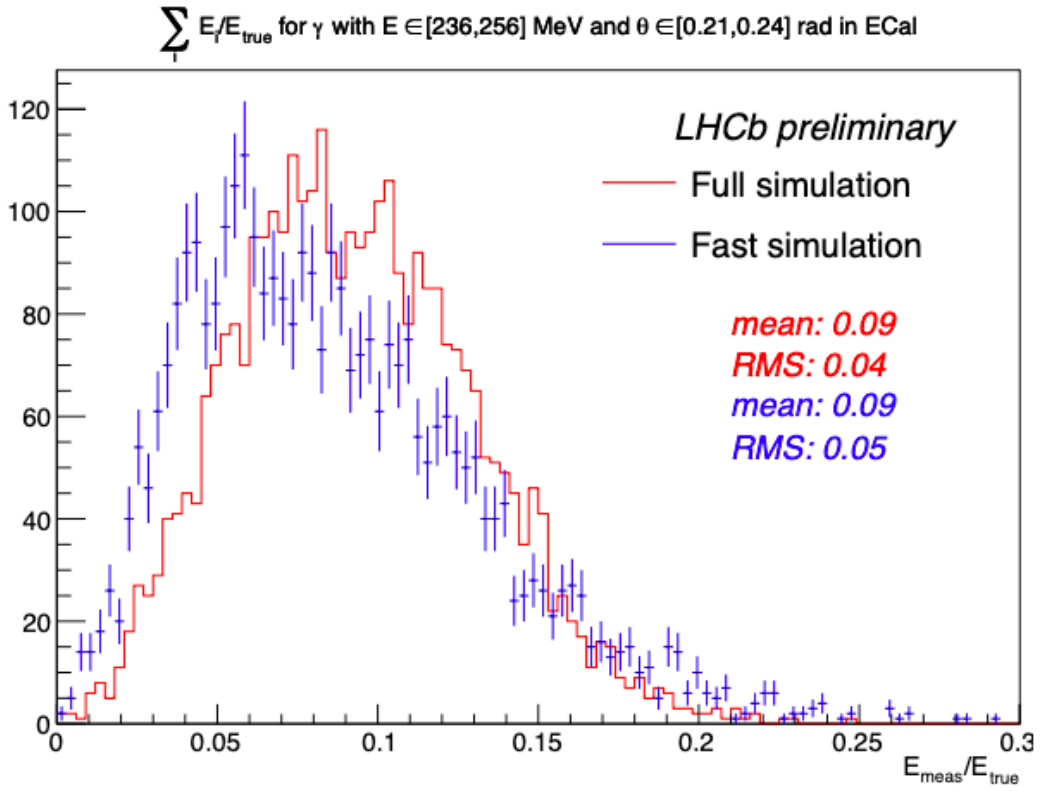


FIG. 2: Layout diagram, and GEANT4 simulation of a single 12.16 GeV electron event in our ACE detector system; in this case liquid nitrogen occupies the interelement spaces.

Ideas for Tabular Methods



Build a library of calorimeter responses to impact particle in corresponding 5D phase space using detailed simulation («frozen showers»).

5D = 3D momentum + 2D coordinate for every particle type.

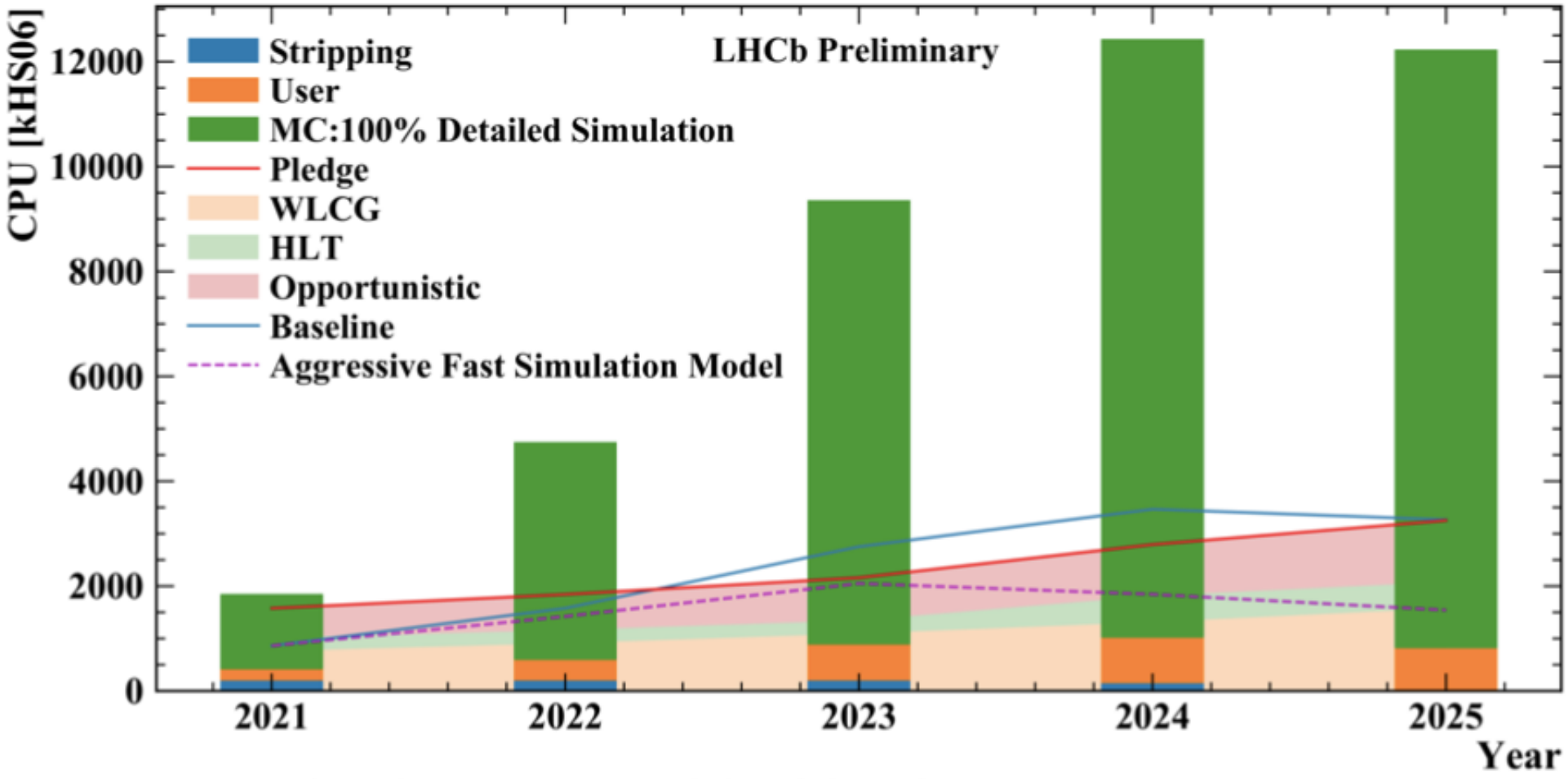
The whole phase space is split into bins, the exact observable is obtained interpolating between the bins.

One can also construct full interpolation (without using bins).

Pros: easy to interpret, quality is controlled by the number of samples.

Cons: curse of dimensionality, memory consumption, full interpolation takes huge efforts.

Upcoming Needs



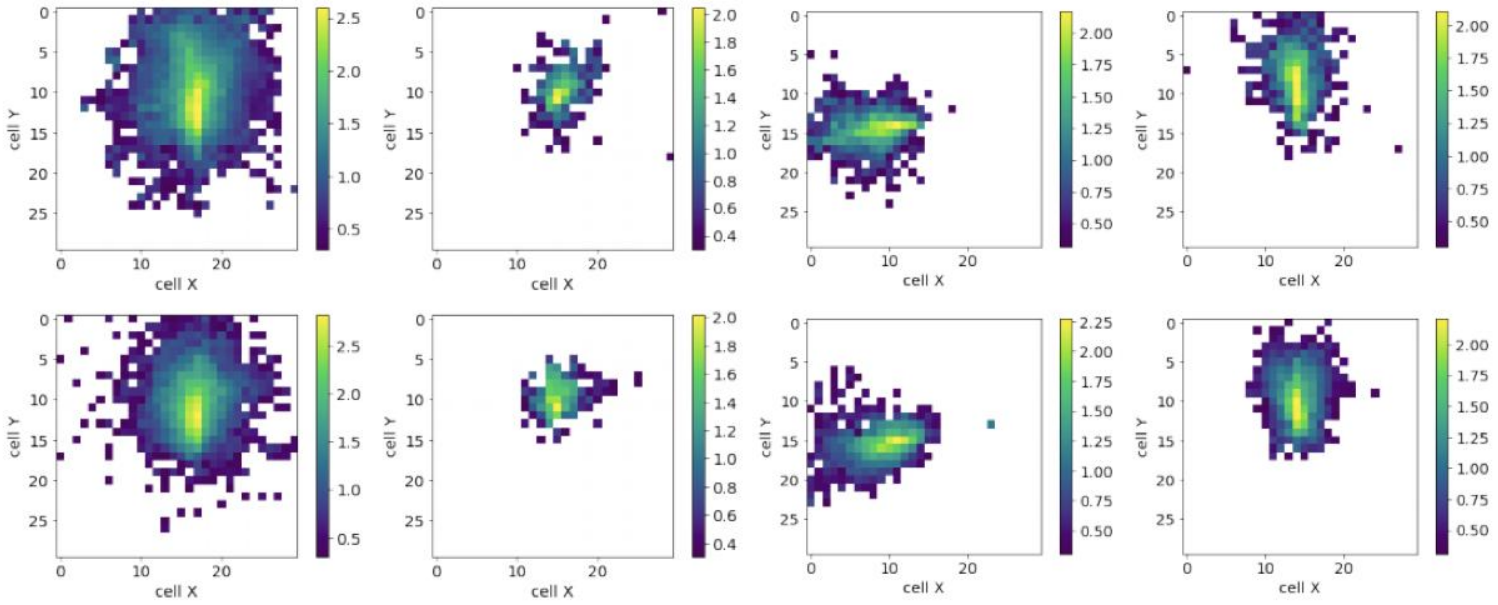
Projected LHCb computing needs breakdown by category

https://indico.cern.ch/event/773049/contributions/3474742/

Generative modeling for HEP

- **Conditional** dependence on incident particle information.

Need to be
- Tunable.
- Robust.
- Fast for sampling.

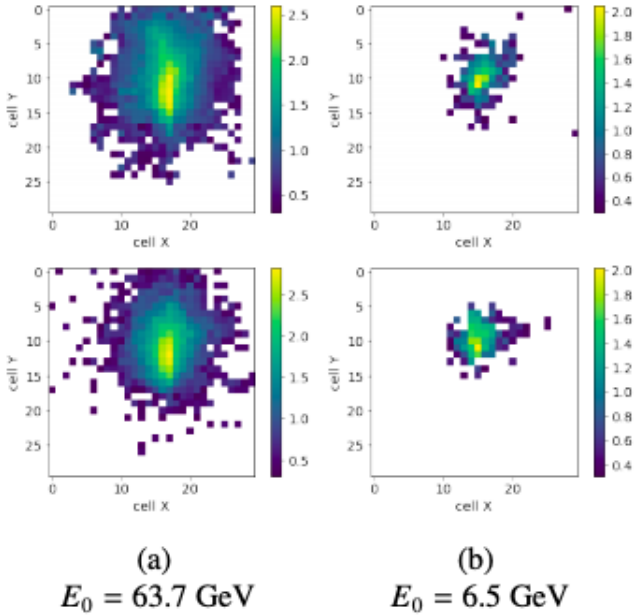


(a) $E_0 = 63.7$ GeV (b) $E_0 = 6.5$ GeV (c) $E_0 = 15.6$ GeV (d) $E_0 = 15.9$ GeV

V. Chekalina et al. *EPJ Web Conf.* 214 (2019) 02034

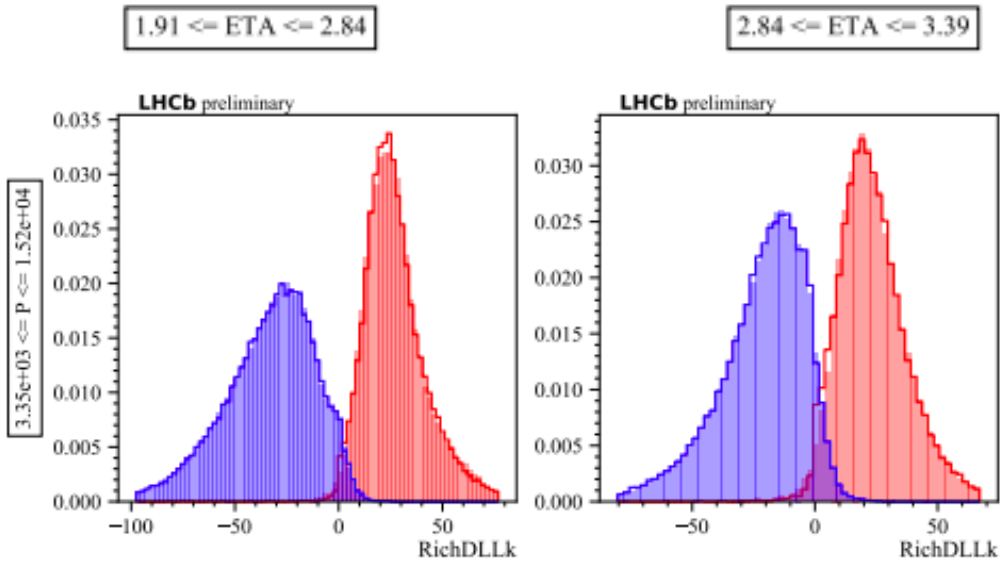
Generative Models

Direct simulation of calorimeter responses



V. Chekalina et al. EPJ WoC: 214, 02034 (2019)

Simulation of reconstruction output for RICH and Muon



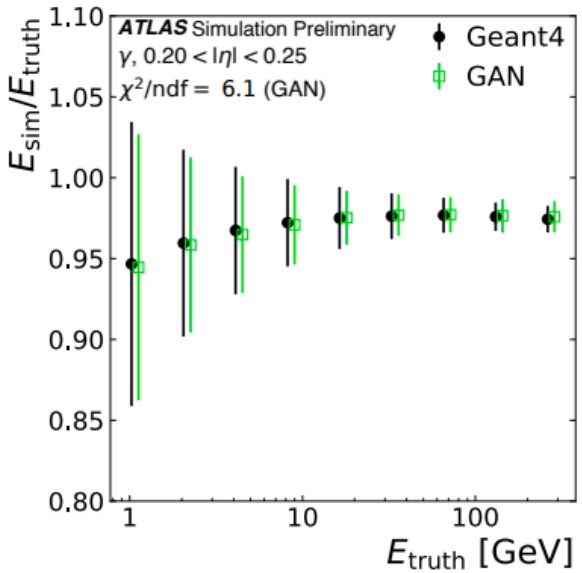
A. Maevskiy et al., ML4PHYS@Neurips 2019

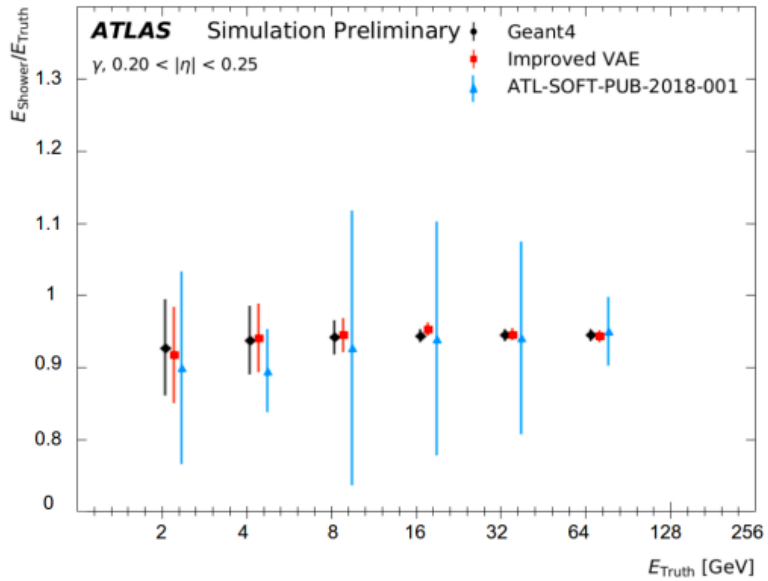
- Reduction of dimensionality of input/outp space can lead to better results.

Generative Models for Fast Simulation

- Many neural based generative description attempted in recent years

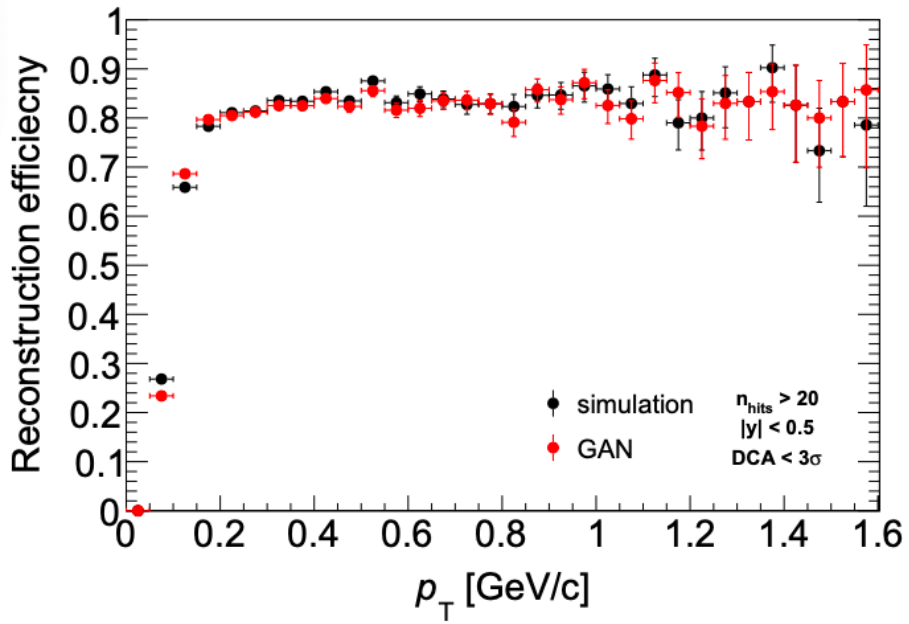
ATLAS: VAE and GAN for Calorimeter





Chapman et al., EPJ Web of Conferences **245**, 02035 (2020)

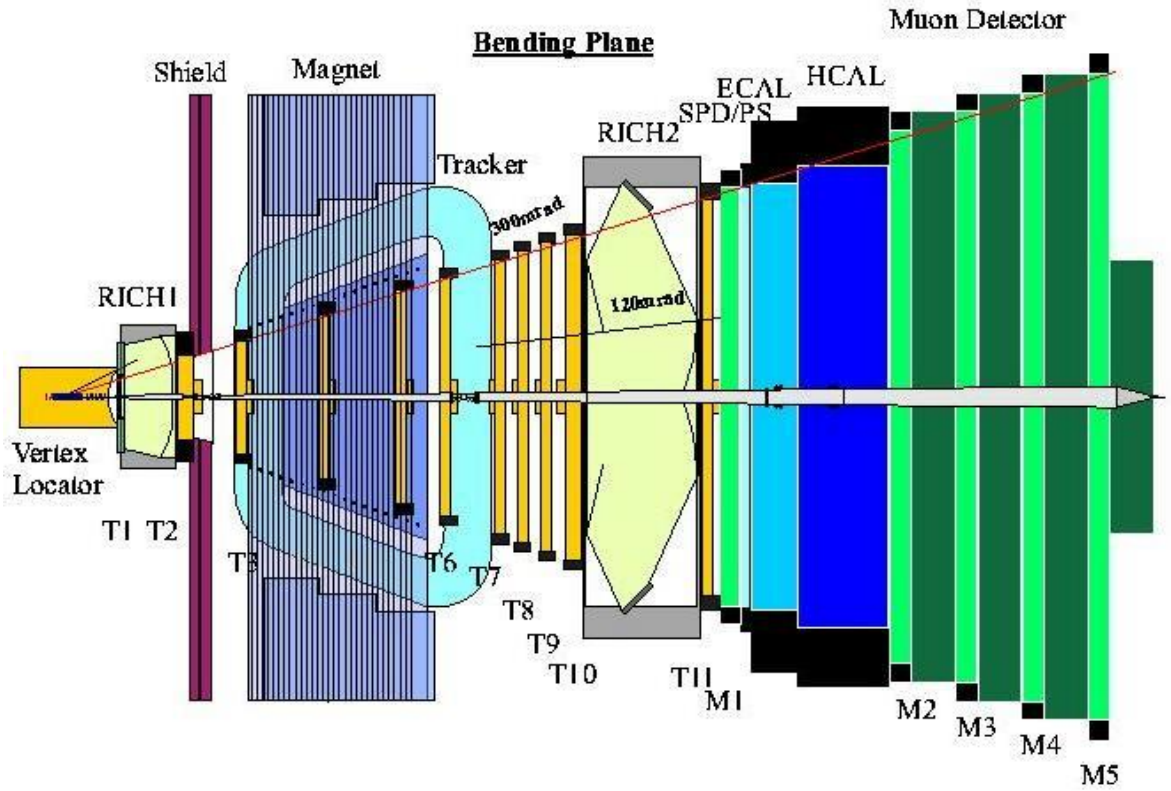
MPD: GAN for TPC



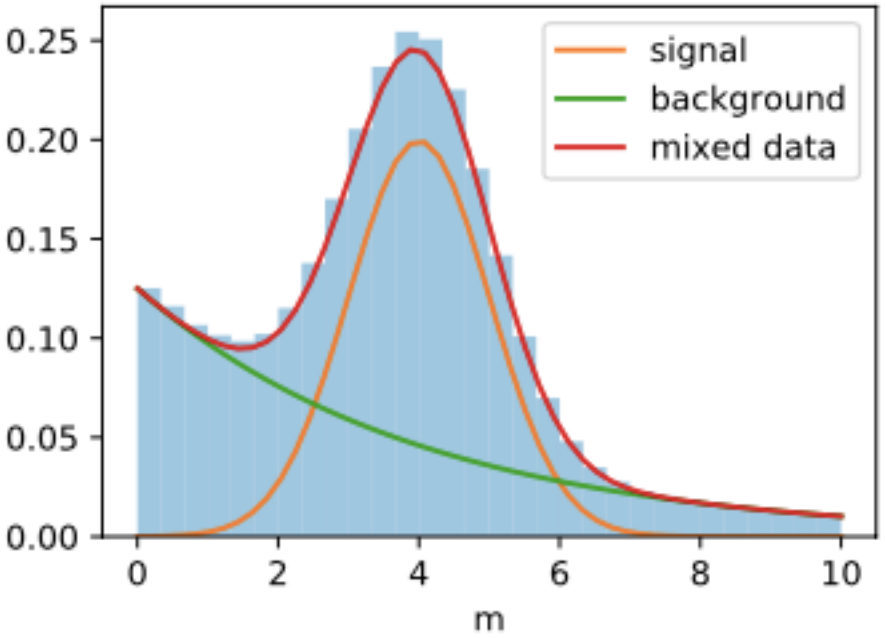
A. Maevskiy et al. Eur.Phys.J.C 81 (2021) 7, 599

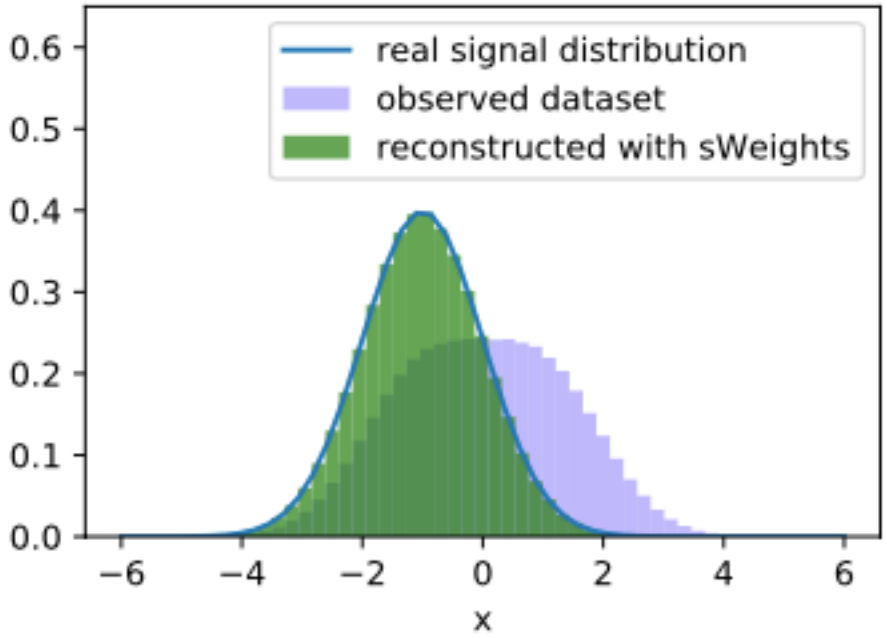
Why it works/should it work?

- Treatment of physics data as pictures.
  - Allows for the use of advanced ML approaches.
- Expressivity of NN solutions.
  - Gives parameterization of sophisticated data.
- Manual decomposition of data.
  - “Expert” approach to model building.



Challenges: Training Samples



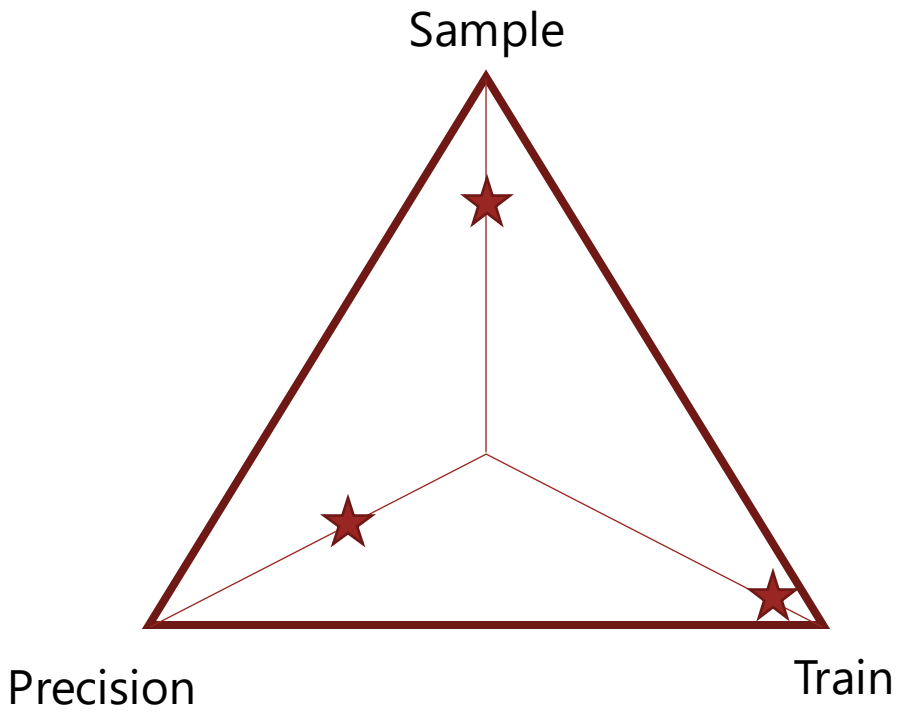


- Use real data sample, but we need to reduce noise from it.
- Model information introduced in the training procedure using maximum likelihood fit.

A. Maevskiy et al., Neurips 2019 Workshop

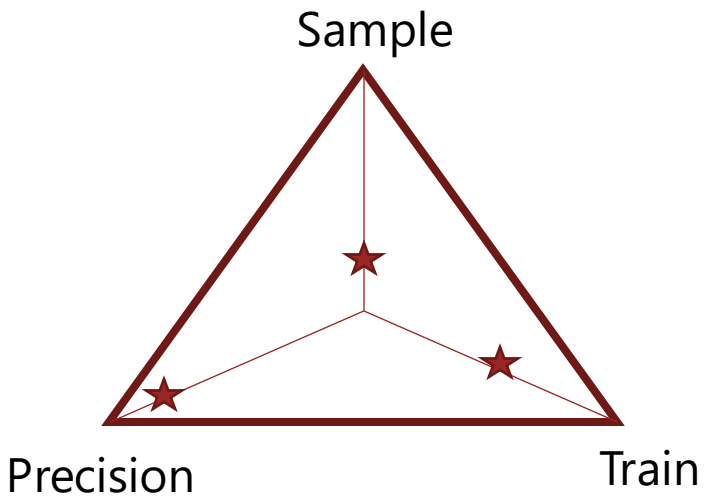
Generative Models Characteristics

- Fast Sampling:
  - much faster than detailed MC;
  - models can get complicated;
  - current simulation speed ~70 ms.
- Very Fast training:
  - retrain can be done very fast;
  - train process still should be periodically controlled;
  - current model trains ~1-2 days using GPU.
- Good Precision:
  - complicated models can be quite precise;
  - precision is controlled by train sample statistics;
  - need to understand influence on the final systematics.

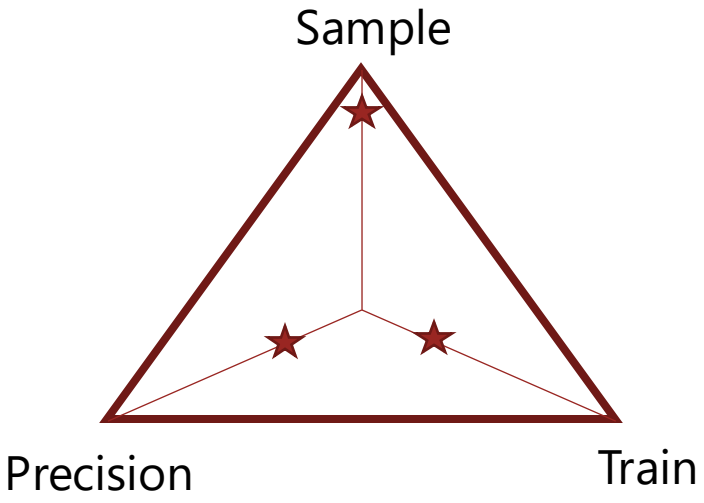


Simulation Picture

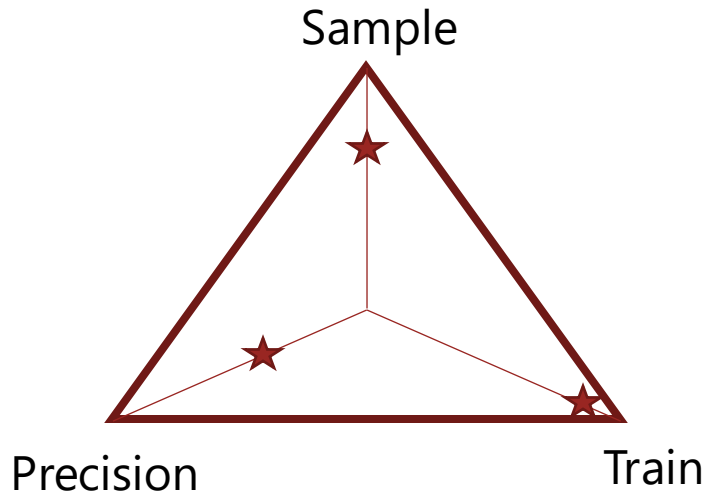
Detailed simulation



Parametric simulation



Machine learning simulation



Each approach has vices and virtues.

A possibility to have easily retrainable model can give several benefits in case of using machine learning.

(*) are my opinion

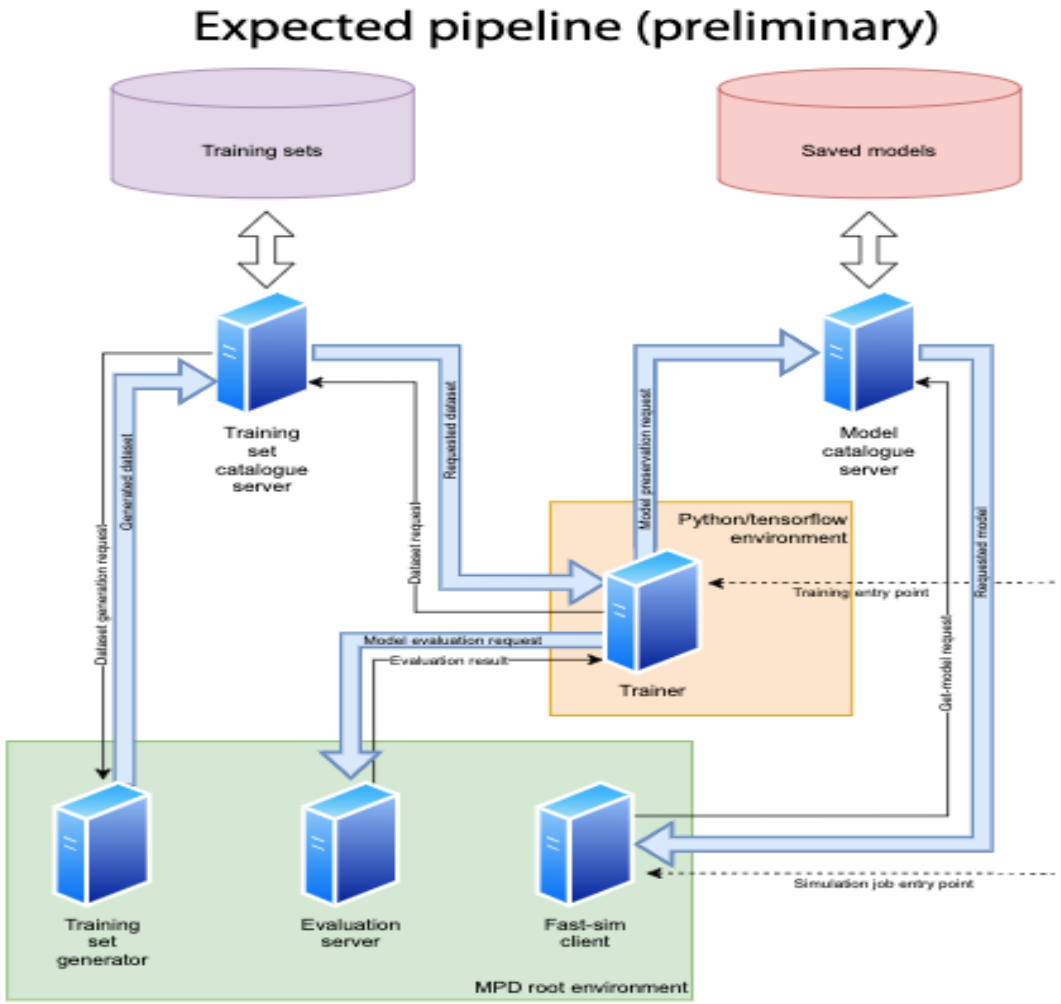
Challenges: Implementations

More challenges:

- **Distilling the generators.**
  Aim: beyond 100ms/event.

- **Testing the generator quality in the limit of small data samples.**
  Aim: on-the-fly algorithms.

- **Implementing pipeline in the online environment (200xNVidia RTX A5000 from LHCb).**
  Aim: Efficient architecture and Scheduling given resources.



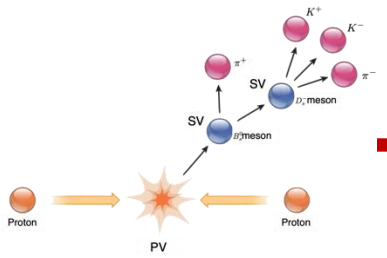
Sukhorosov BSc Diploma

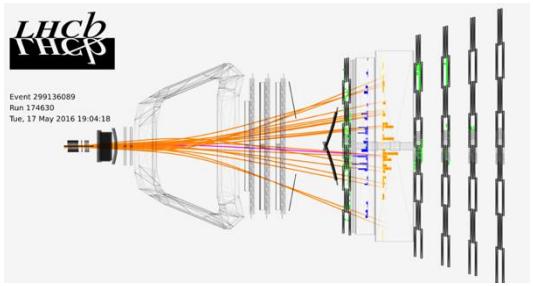
Fast Sampling: take home message

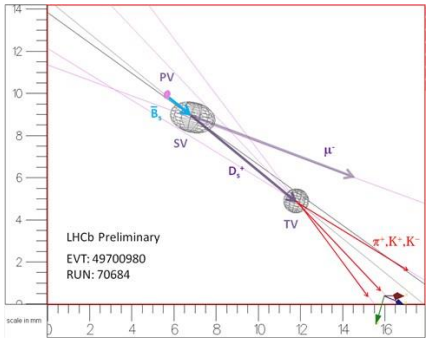
- Machine learning provide a flexible solution but not the fastest.
- Save human time in providing good simulation.
- Good addition to the “classical” approaches.
- Can be retrained online with new samples.

Surrogates

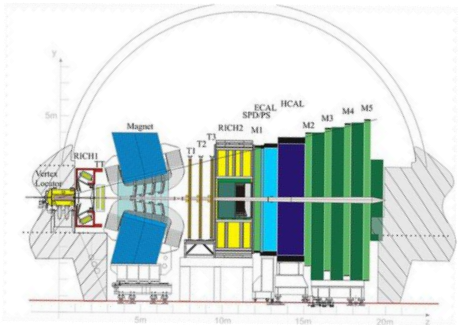
Closer Look







$$\mathcal{L}_{GWS} = \sum_f (\bar{\Psi}_f (i\gamma^\mu \partial\mu - m_f)\Psi_f - eQ_f \bar{\Psi}_f \gamma^\mu \Psi_f A_\mu) +$$
$$+\frac{g}{\sqrt{2}} \sum_i (\bar{a}^i_L \gamma^\mu b^i_L W^+_\mu + \bar{b}^i_L \gamma^\mu a^i_L W^-_\mu) + \frac{g}{2c_w} \sum_f \bar{\Psi}_f \gamma^\mu (I^3_f - 2s^2_w Q_f - I^3_f \gamma_5) \Psi_f Z_\mu +$$
$$-\frac{1}{4}|\partial_\mu A_\nu - \partial_\nu A_\mu - ie(W^-_\mu W^+_\nu - W^+_\mu W^-_\nu)|^2 - \frac{1}{2}|\partial_\mu W^+_\nu - \partial_\nu W^+_\mu +$$
$$-ie(W^+_\mu A_\nu - W^+_\nu A_\mu) + ig' c_w (W^+_\mu Z_\nu - W^+_\nu Z_\mu|^2 +$$
$$-\frac{1}{4}|\partial_\mu Z_\nu - \partial_\nu Z_\mu + ig' c_w (W^-_\mu W^+_\nu - W^+_\mu W^-_\nu)|^2 +$$
$$-\frac{1}{2}M^2_\eta \eta^2 - \frac{gM^2_\eta}{8M_W}\eta^3 - \frac{g'^2 M^2_\eta}{32 M_W}\eta^4 + |M_W W^+_\mu + \frac{g}{2}\eta W^+_\mu|^2 +$$
$$+\frac{1}{2}|\partial_\mu \eta + iM_Z Z_\mu + \frac{ig}{2c_w}\eta Z_\mu|^2 - \sum_f \frac{g}{2}\frac{m_f}{M_W}\bar{\Psi}_f \Psi_f \eta$$



Data Collection

Reconstruction

Hypothesis check

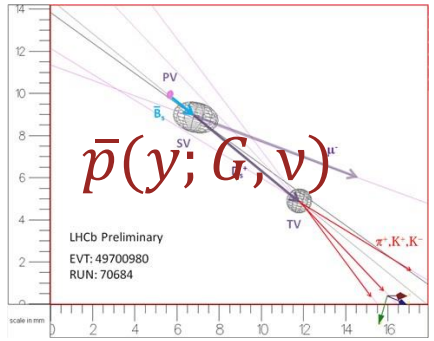
Difficulties

Automate

$L \sim exp(₱_{fix} - ₱)^2 < (p - \bar{p})^2 >$

p
(physics)

$G(x; p, \theta)$
stochastics



- $G(x; p, \theta)$ physics simulator ~1event/minute
- ν chosen manually
- L not differentiable

θ
(detector)

ν
(algorithm)

SIMULATION

Data Reconstruction

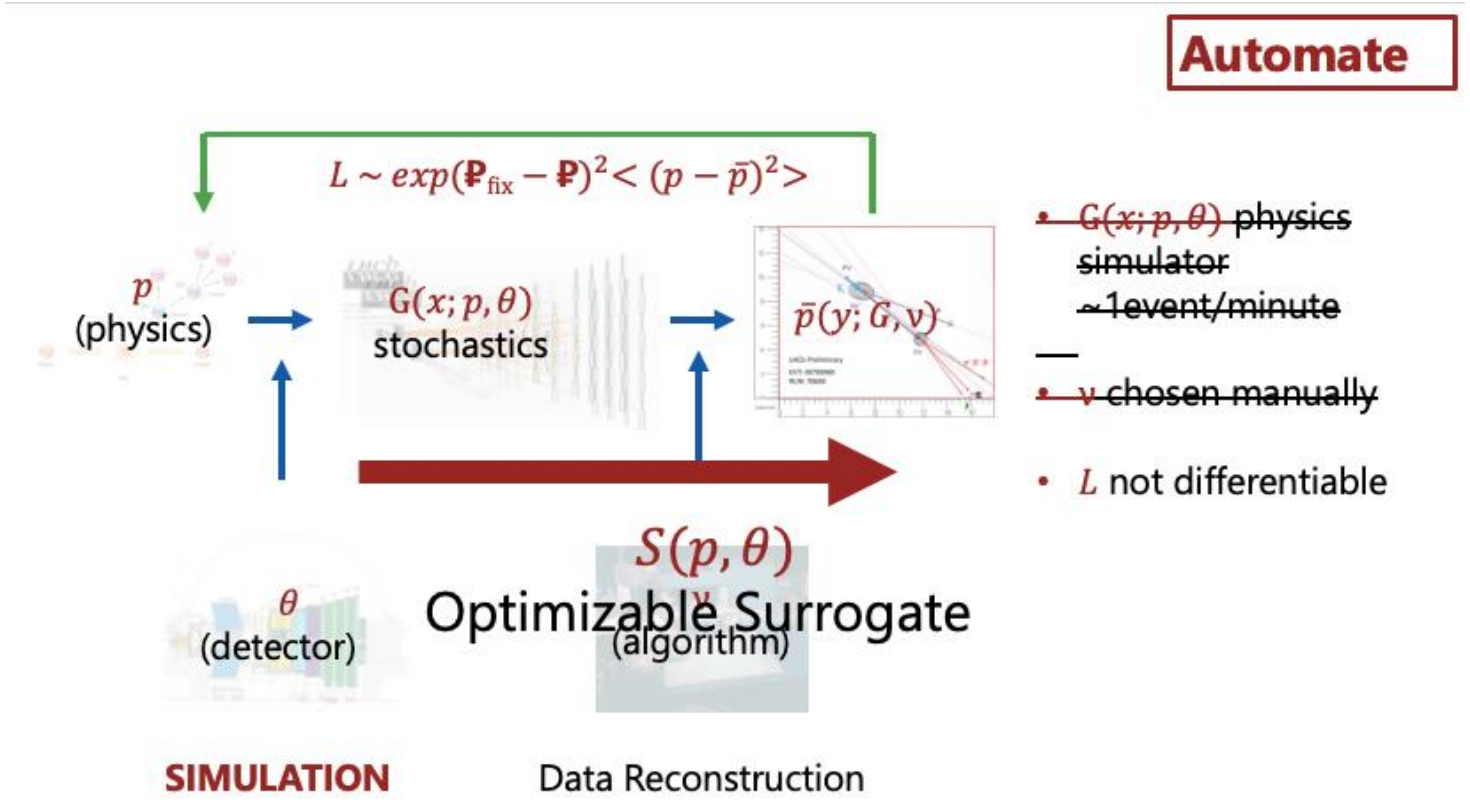
Black box optimization

$$L \sim exp(₽_{\text{fix}} - ₽)^2 < (p - \bar{p})^2 >$$

- The value at any point is known.
- The analytical formula is unknown.
- The time to compute the value is several tens of hours.
- Classical black-box optimization problem:
  - Optimizing someone else's code (there is only a compiled library).
  - Systems described by differential equations (airplane wing shape).

Black Box Solution

- Expert Method
  - Might be wrong
- Random search.
  - Works sometimes.
- Surrogate modeling



“Classic” Surrogate

- Use neural network to connect parameters and outcomes.
- Predict mean behavior.
- Does not take into account variation of distributions depending on the input parameters.

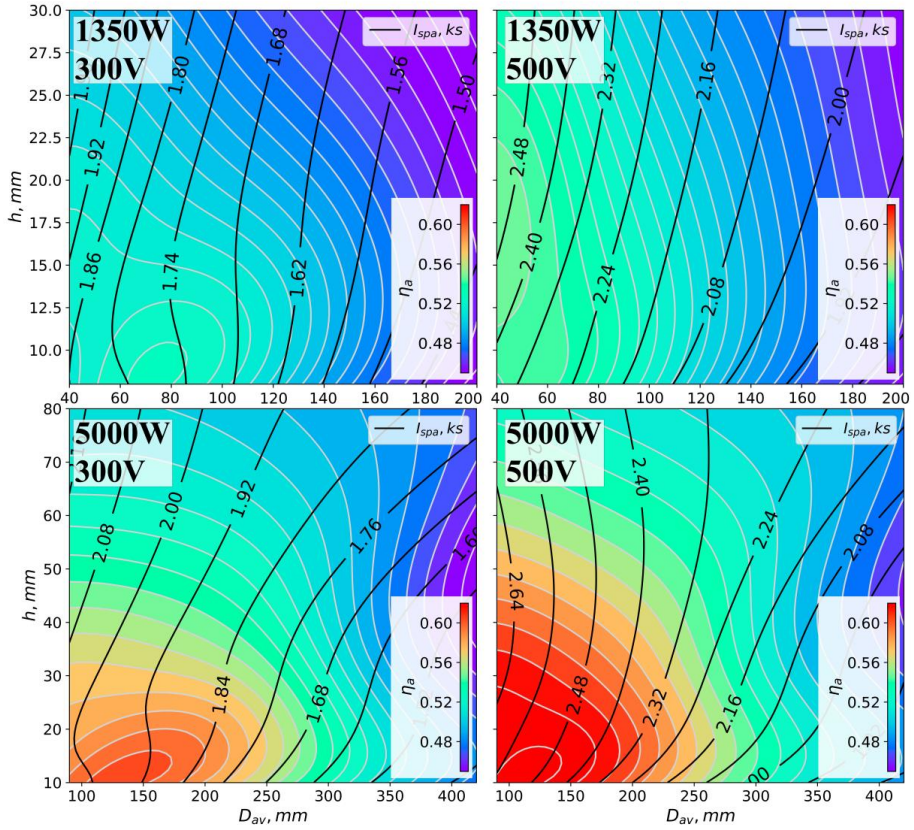
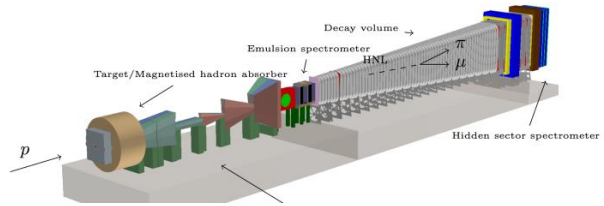
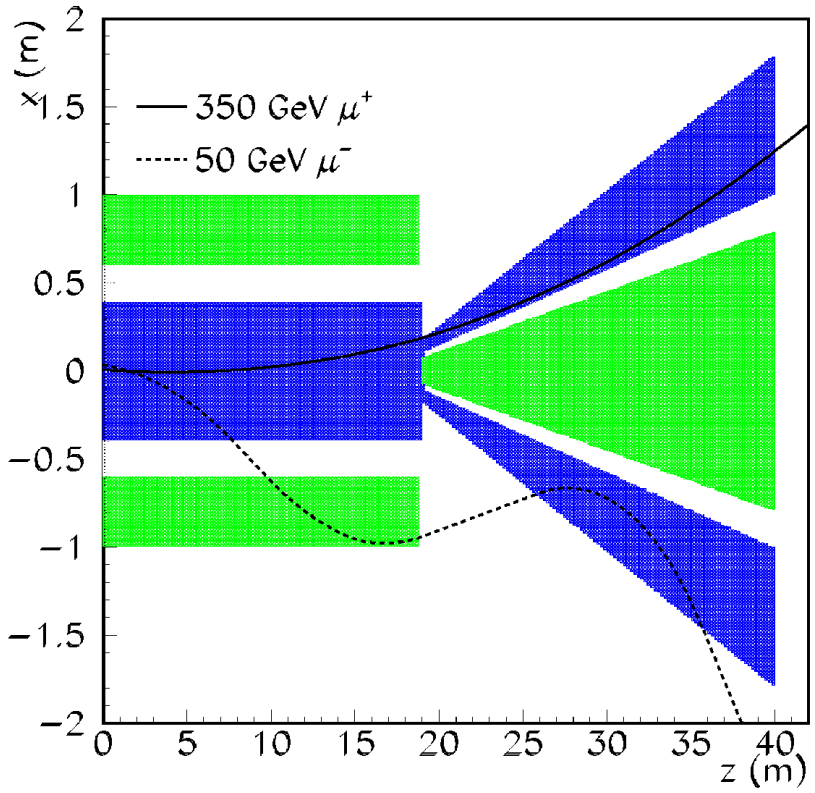


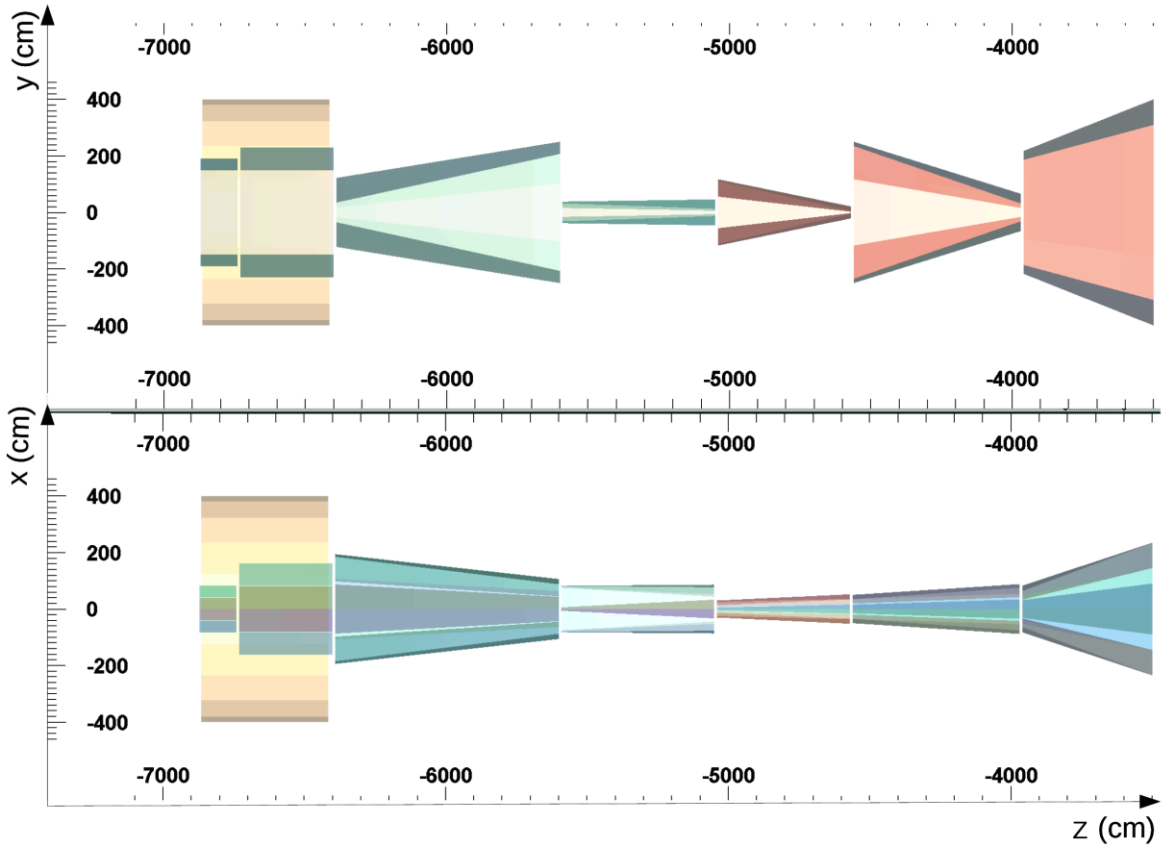
Fig. 14 Predictions from the FNN ensemble of the anode efficiency η_a and anode specific impulse I_{spa} as functions of the discharge channel geometric parameters for the indicated discharge voltages and powers.

Y. Plyashkov et al. On Scaling of Hall-Effect Thrusters Using Neural Nets Journal of Propulsion and Power 2022 38:6, 935-944

SHiP Experiment Design







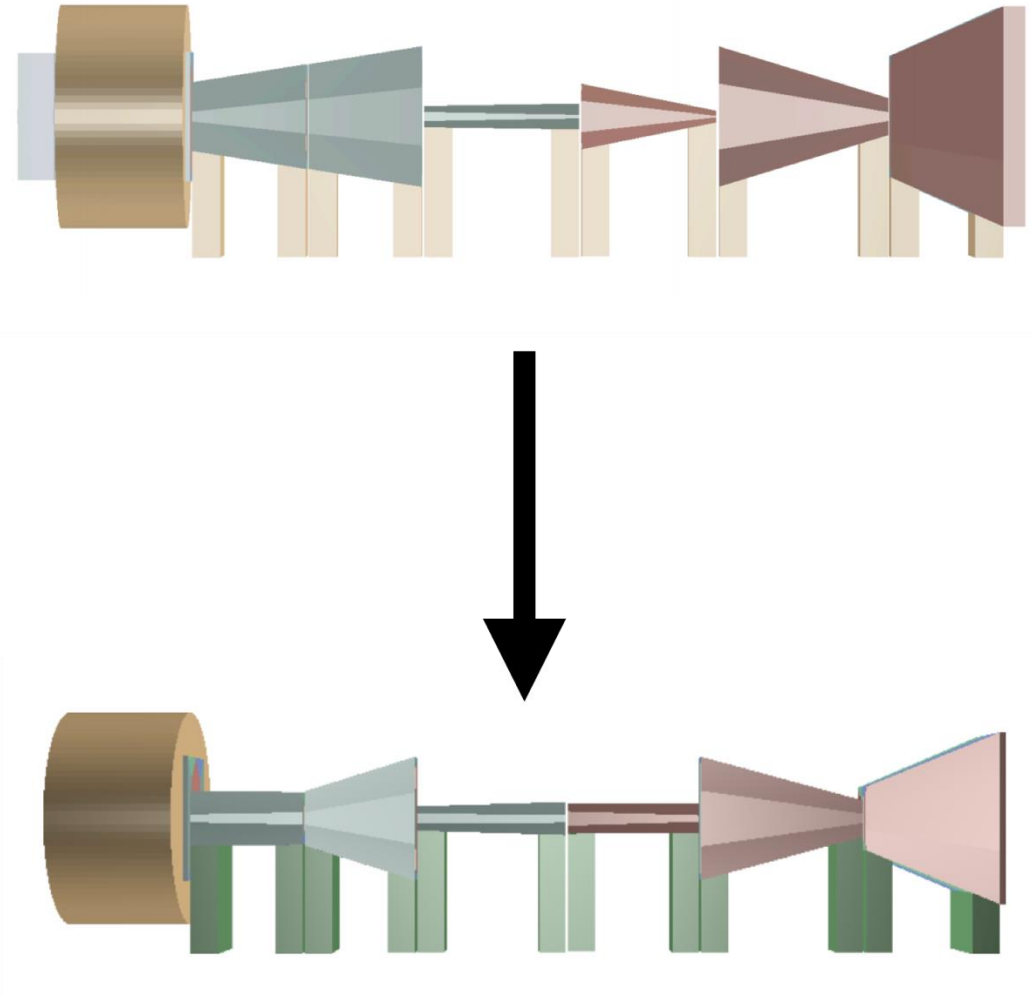
- Absorber shape optimization: background suppression at reasonable cost

Final Optimization

Surrogate modeling using Bayesian Optimization of Gaussian Processes.

Optimization brought 25% cheaper solution.

Currently being tested with engineers.



A. Filatov et al. Journal of Physics: Conference Series. 2017. Vol. 934. P. 1-5

More Opportunities

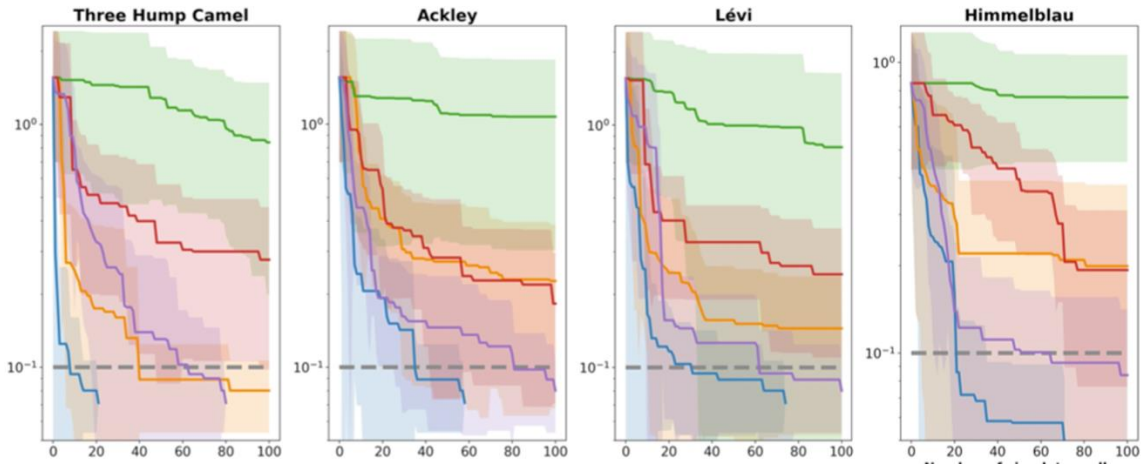
Since we have information about distribution in each particular parameter value.

Algorithm 1 Wasserstein Uncertainty Global Optimisation (WU-GO)

Input: Ground truths $\mathcal{M}$, generator G, grid $\tilde{\Theta}$, parameter κ
Output: Optimal configuration $\hat{\theta}^*$

---

while stopping criteria are not met **do**
  Fit G on $\mathcal{M}$
  Approximate f: $\hat{f}(\theta) = \mathbb{E}[G(\theta)] \approx \frac{1}{n}\sum_{j=1}^{n} x_j, x_j \sim G(\theta)$
  Estimate $\sigma_{\mathbb{W}}$: $\hat{\sigma}_{\mathbb{W}}(\theta) = \min_{\mu \in \mathcal{M}} D(\mu, G(\theta))$
  Predict $\hat{\theta} = \arg\min_{\theta \in \tilde{\Theta}} \{\hat{f}(\theta) - \kappa \cdot \hat{\sigma}_{\mathbb{W}}(\theta)\}$
  Call simulator for $\hat{\theta}$: $\hat{\mu}$
  $\mathcal{M} = \mathcal{M} \cup \{\hat{\mu}\}$
end while



We can use it to estimate the most interesting point for the next step of optimization.

This is very useful for non diff black box optimization.

T. Ramazyan et al. ECAI-2024

Surrogates' conclusions

- Optimization of large setups requires simultaneous approximate solution of forward and inverse problems.
  - Many challenges in both parts: speed-up, implementation, tail control of generative models.
- Final stage of the optimization brings the need of precise surrogate modeling.
- A complete optimization cycle brings in significant reduction in costs (and thus efficiencies) but requires several fundamental questions to be solved.

Overall conclusions

- Generative modeling usage, while being effective in some applications, remains a challenge real-world applications.
- The main challenges are:
  - speed
  - implementation
  - uncertainty
  - data hunger

In next several years, one can expect significant number of results in the applied generative modeling field.