

Survey

A survey on generative AI for detector effects unfolding in particle and nuclear physics

Tareq Alghamdi^{1,2,†}, Tommaso Vittorini^{3,†}, Jitao Xu¹, Marco Battaglieri³, Derek I. Glazier⁴, Glòria Montaña^{5,6}, Giorgio Foti^{7,8}, Alessandro Pilloni^{7,8}, Nobuo Sato⁵ and Yaohang Li^{1,*}

¹ Department of Computer Science, Old Dominion University, Norfolk, VA 23529, USA

² Department of Computer Science, Al-Baha University, Al-Baha, Alaqiq 65779, Saudi Arabi

³ INFN sezione di Genova, I-16146 Genova, Italy

⁴ SUPA, School of Physics and Astronomy, University of Glasgow, Glasgow G12 8QQ, United Kingdom

⁵ Thomas Jefferson National Accelerator Facility, Newport News, VA 23606, USA

⁶ Department de Física Quàntica i Astrofísica and ICCUB, Universitat de Barcelona, E-08028 Barcelona, Spain

⁷ Università degli studi di Messina, I-98166 Messina, Italy

⁸ INFN Sezione di Catania, I-95123 Catania, Italy

[†] These authors contributed equally to this work.

* **Correspondence:** Email: yaohang@cs.odu.edu.

Academic Editor: Pasi Fränti

Abstract: In particle and nuclear physics, “detector effects unfolding” can be viewed as a high-dimensional inverse problem whose goal is to recover the true event distributions from observed experimental data corrupted by detector-induced distortions. Recent advances in generative AI have positioned data-driven and machine learning-based approaches as powerful alternatives to traditional unfolding techniques, offering superior scalability to high-dimensional data, capability of learning complex detector responses, and the ability to operate directly at the event level. We survey state-of-the-art generative AI-based models for detector folding and unfolding. We review existing architectures and training strategies, and highlight recent methodological advances and open challenges. Through a detailed discussion of latent space representations, uncertainty quantification, physics insights, and background removals, we demonstrate the potential to significantly advance the field.

Keywords: generative AI; detector effects; unfolding; particle physics; nuclear physics; uncertainty quantification; inverse problems

1. Introduction

The particle physics approach to probing the fundamental laws of nature relies on accelerators. These machines propel particles to extremely high energies and let them collide with each other or with stationary targets. Examining the outcome of these collisions through sophisticated detectors, we can test our knowledge of nature at the femtoscale, comparing the results of these experiments with theoretical predictions.

Prominent facilities driving this field include the large hadron collider (LHC) at CERN, which focuses on proton-proton collisions; the continuous electron beam accelerator facility (CEBAF) at Jefferson Lab, specializing in polarized electron-nucleon scattering; and the relativistic heavy ion collider (RHIC) at Brookhaven National Laboratory, which investigates proton and nuclear collisions. Looking ahead, the upcoming electron-ion collider (EIC) aims to push the boundaries of precision in probing nuclear matter.

These experiments produce a huge volume of data, yet the insights derived from this data are constrained by the limitations of particle detectors [56]. Detectors, although highly sophisticated, are not perfect [16]. They introduce effects such as smearing, acceptance, and inefficiencies, which distort the true physics processes that we are trying to understand. To address this important challenge, a common strategy is to rely on *unfolding*, which aims to reverse these detector effects and reconstruct the original particle interactions. Conversely, *folding* involves incorporating these effects into theoretical predictions to simulate how the detector would perceive them, which enables a comparison between experimental data and synthetic data generated by theoretical models. Figure 1 illustrates a schematic view of the folding and unfolding procedures. Unfolding is important because detector distortions can shift measured distributions, obscure narrow physics structures, and bias downstream parameter extraction. From an AI perspective, it is also a high-dimensional inverse problem with incomplete observations, biased sampling, physical constraints, and unavailability of ground truth.

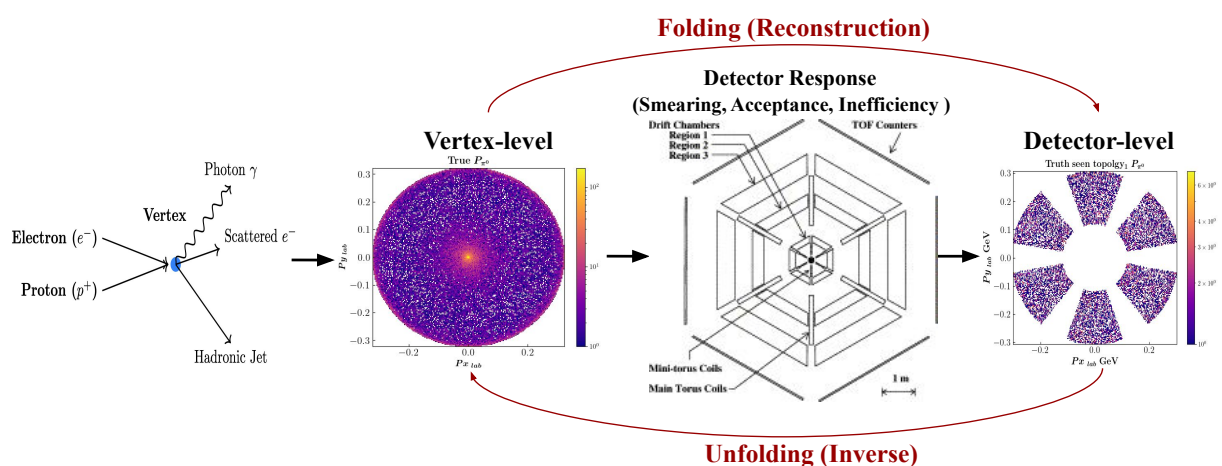


Figure 1. A schematic view of the folding and unfolding procedures.

Traditional approaches to detector folding and unfolding often rely on classical statistical inference and handcrafted modeling assumptions about the underlying physics. Although these methods have achieved notable success [17,27], they exhibit several fundamental limitations from a modern machine learning (ML) perspective. First, many traditional algorithms struggle to scale up to the high-dimensional data in modern particle physics experiments such as those in LHC and CEBAF. Second, commonly used techniques, for example, Bayesian unfolding, frequently rely on simplified analytical models or linear assumptions about the detector response, which may not represent the complexity of the underlying physics. Third, many traditional methods are specifically designed for particular types of detector effects or experimental setups, which offer limited generalization. Finally, an important drawback is that histogram-based analysis relies on data binning, which not only requires manual fine-tuning but also can introduce biases and lead to a loss of valuable information.

Recent advances in generative artificial intelligence have introduced a fundamentally different, data-driven paradigm for addressing detector folding and unfolding [18,24]. Generative AI-based models, such as *generative adversarial networks* (GANs) [39], *variational autoencoders* (VAEs) [45], *normalizing flows* (NFs) [59], and *diffusion models* (DMs) [41], offer several key advantages over traditional methods for unfolding detector effects. Powered by deep neural networks, generative models can effectively learn complex relationships in high-dimensional space, particularly suited for unfolding events with many measured particles. Learning directly from data, generative models can learn the complex, nonlinear response of detectors without relying on restricted analytical assumptions. When uncertainty quantification (UQ) mechanisms are incorporated, generative models can achieve precise statistical analysis in the unfolded results. Moreover, by performing event-level unfolding and avoiding binning, generative models can preserve the full content of the data without losing any information. Furthermore, generative AI-based methods can be designed to unfold multiple observables across different experiments, capturing cross-correlations and improving inference consistency and accuracy.

Despite their promise, generative AI-based unfolding methods remain at an early stage of development, and careful validation, robustness analysis, and interpretability studies are essential for their reliable deployment. This paper presents a comprehensive survey of generative AI approaches for detector effects unfolding by reviewing state-of-the-art generative models designed to address the acceptance, smearing, and inefficiency effects introduced by detectors and highlighting their advantages, challenges, and applications across different experiments. In addition, we discuss key methodological themes, including latent space analysis, background removals, uncertainty quantification, the extraction of physics from generative models, and multiple topology unfolding. Through this survey, we aim to provide a clear computational roadmap for researchers exploring generative AI solutions to this critical challenge in high-energy physics data analysis.

2. Detector effects

In any particle experimental setup, the observed data is influenced not only by the underlying physical processes but also by the interaction of particles with the detector. These interactions introduce various distortions that must be carefully accounted for in order to extract meaningful physics results. Figure 1 illustrates how detector responses such as smearing, acceptance, and inefficiency distort true vertex-level data into observed detector-level data. Detector effects can impact the accuracy, precision,

and completeness of measured distributions, often necessitating complex correction procedures.

Broadly, these effects can be categorized into *smearing*, *acceptance*, and *inefficiency*. Smearing arises because of the finite resolution of the detector, leading to distortions of the distributions of measured quantities, such as energy and momentum. This can cause interesting physics effects, which could be shown as narrow peaks, to be completely washed away. Acceptance effects result from the detector's inability to cover the full space allowed by the reaction due to geometric limitations. This leads to a loss of entire sections of the observable's distributions, making data interpretability significantly harder. Inefficiency, on the other hand, is related to the detector's failure to register certain particles, introducing missing data that can bias analyses.

Addressing these effects is crucial for reliable data interpretation. Various approaches, including simulation-based corrections and unfolding techniques, are employed to mitigate detector distortions and recover the true underlying distributions. A particular type of these distortions, based on the CEBAF Large Acceptance Spectrometer (CLAS) detector, can be observed in Figure 2.

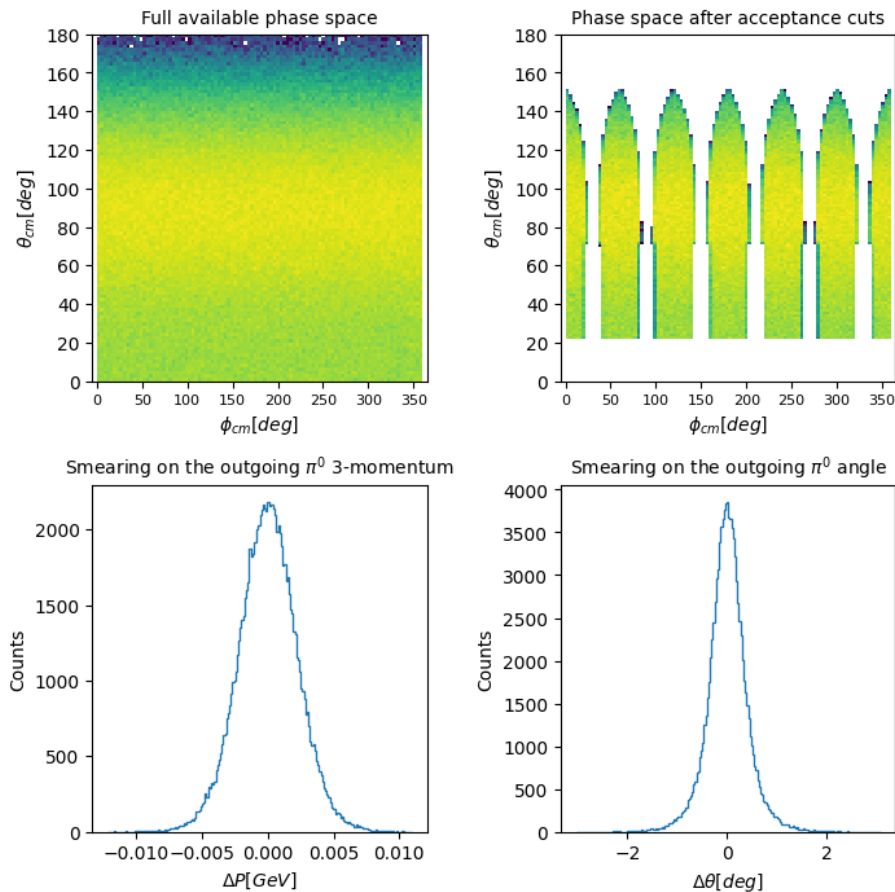


Figure 2. CLAS single- π^0 photoproduction sample. Top: available phase space in terms of one independent variable of single π^0 photoproduction before (left) and after (right) applying cuts similar to the ones present in the CLAS detector. Bottom: smearing distortion on the 3-momentum (left) and scattering angle (right) of the outgoing π^0 in the center-of-mass frame.

3. Folding detector effects

In the context of folding detector effects, various generative AI models have been explored, each tailored to specific aspects of detector simulation, including calorimetry, tracking, and Cherenkov detectors. Among these, GAN-based approaches have been widely utilized to model detector responses, followed by VAE-, NF-, and DM-based approaches. Table 1 provides a comparative overview of these models. Because the surveyed studies use different datasets, detector configurations, observables, training procedures, and model hyperparameters, Tables 1 and 2 should be interpreted as qualitative summaries of reported trends rather than controlled benchmark comparisons. Shared datasets, standardized metrics, and consistent hyperparameter reporting remain important needs for future comparisons of generative models for detector effects unfolding.

Table 1. Comparison of generative models for handling detector folding. Training stability refers to how reliably and predictably the generative model converges during optimization without divergence, oscillations, or collapse. Scalability describes how efficiently a model's training and inference performance grows with dataset size and dimensionality. Inference speed is the time required to compute conditional outputs from a trained model. Fidelity refers to how accurately generated samples match the true data distribution. Expressivity is a generative model's theoretical and practical capacity to represent complex, high-dimensional probability distributions.

Model	Training Stability	Scalability	Inference Speed	Fidelity	Expressivity
DMs	High	Moderate-High	Slow(many denoising steps); Fast(with distillation)	High	High
GANs	Unstable	High	Very Fast (single forward pass)	High (but prone to mode collapse)	Moderate
VAEs	Moderate-High	High	Very Fast (single pass)	Moderate	Moderate
NFs	Moderate (depends on architecture constraints)	Moderate	Fast-Medium (direct sampling, but deep layers slow down)	High	Moderate

Table 2. Comparison of unfolding methods in particle physics, categorized by approach. Correlations refer to the interdependencies between observables, events, and uncertainties that must be preserved to recover physically and statistically consistent truth-level distributions. Binning sensitivity refers to how strongly the unfolded results depend on the choice of histogram binning used to discretize continuous variables.

Unfolding Method	Scalability	Correlations	Binning Sensitivity	Interpretability	UQ
Tikhonov	Low	Limited	High (bin-dependent)	High	Well-defined (regularization)
Regularization/Bayesian					
GAN-based Unfolding	Moderate-High	Moderate-High	Low	Low (black-box)	Moderate (requires ensembles)
INN	Moderate-High	High-Very High	Low	High-Very High	Intrinsic
Diffusion models	High	High-Very High	None-Low	Low-Moderate	High
Moment Unfolding	High	High	None (moment-based)	High	Strong via statistical methods

3.1. GAN-based approaches

A range of generative adversarial network (GAN) architectures have been developed to model detector responses, particularly for calorimeters and tracking systems. Early work such as lesion-aware generative adversarial networks (LAGAN) [30] demonstrated the feasibility of learning calorimeter responses by reconstructing two-dimensional jet images using locally connected layers, enabling the model to capture location-dependent spatial patterns. Building on this idea, CaloGAN [53] and its successor CaloShowerGAN [37] adopted conditional GAN frameworks to generate layerwise calorimeter images conditioned on the incident particle energy, allowing the models to generalize across a wide energy range. To further improve sample quality and mitigate performance degradation in low-energy regions, ensemble-based extensions leveraging adaGAN have been proposed [43].

For tracking detectors, architectures have evolved to handle ultra-high granularity data. The intra-event aware generative adversarial network (IEA-GAN) [14] integrates adversarial training with transformer-based architectures and self-supervised objectives, enabling the model to capture long-range dependencies and event-level correlations across complex detector geometries. GRAPH-GAN [44] employs graph-based learning to explicitly model the topological structure inherent in tracking data, leading to the improved simulation of sparse detector events.

Beyond detector-specific simulations, GANs have been applied to event-level processes where smearing is introduced before unfolding. For instance, recent works such as [4, 7] used GANs to learn event-level detector effects, applying smearing to simulated physics processes before reconstructing underlying distributions. These approaches commonly adopt stabilized adversarial objectives, such as least-squares GANs (LSGANs) and conditional formulations (cGANs) to improve training robustness and provide finer control over the generated event distributions, highlighting the flexibility of GAN-based models for both detector- and event-level simulations.

From a computational perspective, GAN-based approaches offer several advantages over traditional Monte Carlo-based simulation pipelines. Once trained, GANs enable orders-of-magnitude faster inference, making them attractive for large-scale data generation and rapid iteration [20]. Their expressive architectures allow them to model complex, high-dimensional phase spaces and capture nonlinear correlations that are difficult to encode with handcrafted simulation models. Moreover, GANs vastly reduce the need for extensive presimulated datasets by encoding detector responses in their network weights [26].

Despite the above advantages, GAN-based methods face well-known challenges such as training instability, mode collapse, and the need for domain-specific adaptations [5]. Addressing these issues often requires architecture regularization and careful incorporation of physical constraints [23]. Furthermore, [3] provides a comprehensive review of related advances in generative models for physics event generations.

3.2. Variational autoencoders (VAEs)

VAEs have been widely adopted in detector simulation, especially for modeling calorimeter responses and Cherenkov detector signals. Although standard VAEs are known to struggle with accurately approximating complex posteriors, several architectural and training enhancements, such as trainable priors with noise generation to encode embeddings in latent space [32], have been proposed to mitigate these limitations. Regularization techniques are added to promote diversity and mitigate

mode collapse, ensuring stability through similarity matrix-based regularization.

Several VAE-based architectures have demonstrated improved performance in calorimeter shower reconstruction. The bottleneck information bottleneck autoencoder (BIB-AE) [21] model improves shower reconstruction by conditioning the latent space on incident photon energy, using a post-processing module to balance hit energy spectrum accuracy and shower shape fidelity. The calorimeter manifold autoencoder (CaloMan) [28] reduces the dimensionality of calorimeter data via latent space learning, followed by density estimation to enable efficient sampling.

For Cherenkov detectors, hybrid conditional VAE frameworks have been proposed to better capture complex detector responses. For example, [31] combines a conditional VAE with an information maximizing VAE (infoVAE) [64], using Bayesian optimization to select latent dimensionality and conditioning reconstruction on particle kinematics, leading to improved sample fidelity. More broadly, incorporating NFs as learnable priors within variational inference [59] enhances posterior expressivity and sampling efficiency for grid-based Water Cherenkov detector data.

To address sparse and irregular detector geometries, graph-structured VAEs have been introduced. The graph VAE (GVAE) [40] leverages graph-based representations to improve modeling of sparse collision events, though with limited controllability in sampling. Related variants, such as denoising VAEs (DVAE) [2] and bicephalous convolutional autoencoder++ (BCAE++) [42], further demonstrate the effectiveness of tailored VAE architectures for sparse, high-dimensional detector data.

More recently, the Yoneda variational autoencoder (YonedaVAE) [14] has been proposed for ultra-high granularity detector simulation, introducing context extrapolation, zero-shot generalization, and variable-cardinality handling; it marks a step toward scalable generative modeling for next-generation detector systems.

3.3. *NF-based models*

Normalizing flows (NFs) provide a principled and expressive framework for modeling detector effects by learning exact, invertible mappings between simple base distributions and complex data distributions. Initially introduced to improve posterior approximation in variational inference [59], NFs have since been adopted as standalone generative models for detector simulation. CaloFlow [49] demonstrated early success by applying NFs to calorimeter data using masked autoencoder for distribution estimation (MADE) blocks and rational quadratic spline transformations. CaloFlow II [48] further improved practicality through knowledge distillation, transferring density estimates from a teacher model to a faster inverse autoregressive flow (IAF) student, significantly accelerating inference. More recently, SuperCalo [54] employed flow-based super-resolution to map coarse calorimeter showers to fine-grained representations, enabling efficient modeling across detector resolutions.

Compared to GANs and VAEs, a key advantage of NFs is their ability to compute exact likelihoods for generated events, facilitating statistically rigorous analysis of detector-induced distortions and uncertainty-aware inference.

3.4. *DM-based models*

Diffusion models (DMs) have recently emerged as a robust and expressive framework for simulating detector effects, offering improved training stability and strong performance in high-dimensional settings. CaloClouds II [22] introduced a geometry-agnostic diffusion approach for highly granular

calorimeter simulation, leveraging continuous-time score-based modeling and consistency distillation to enable efficient, single-step sampling with significant speedups over traditional simulations. CaloDiffusion [9] employs denoising diffusion with 3D cylindrical convolutions to exploit detector symmetries along with a geometry-latent mapping layer to handle irregular detector layouts, achieving near-simulation-level fidelity. Beyond calorimetry, diffusion models have been applied to fast simulation of reconstructed particles [46], using set-conditional generation to emulate complex jet-level particle spectra within unified simulation-reconstruction pipelines.

Unlike other generative folding methods, DMs iteratively refine samples through denoising, ensuring broader coverage of phase space. This makes them well-suited for representing diverse reactions across a wide kinematic range. Moreover, the simplified training process, based on minimizing a single loss function, mitigates issues such as mode collapse and training instability, leading to more robust and stable performance compared to the adversarial nature in GANs or variational inference in VAEs.

4. Unfolding detector effects

This section mirrors the model-based organization used for folding so that one can compare how each family of generative methods is adapted to the inverse problem. In recent years, deep learning techniques have significantly advanced unfolding methods in particle physics, addressing limitations of traditional approaches, which we can see a summary of in Table 2. Early methods such as Tikhonov regularization and iterative Bayesian unfolding, widely used in 2010, provided robust solutions but struggled with scalability, maintaining correlations, and sensitivity to binning choices [17].

4.1. GAN-based unfolding

The application of GANs to unfolding detector effects emerged in 2018, enabling data-driven corrections [29]. Building on this, mean squared error GAN (MSGAN), a modified GAN architecture, was proposed in [53] to model detector effects and perform unfolding. Although it performed well across various distributions and outperformed traditional methods such as Bayesian and singular value decomposition (SVD) unfolding, MSGAN was computationally intensive and highly dependent on the quality of the training dataset. Another GAN-based approach, fully conditional GANs (FCGANs), was introduced in [15] to invert detector simulations, reconstructing parton-level events with greater flexibility and precision in multidimensional scenarios. Recent approaches have extended generative unfolding techniques beyond detector-level corrections to full event-level reconstruction. Works such as [4, 7] explored the use of a framework composed of two GANs, an inner GAN to model detector smearing effects at the event level and an outer GAN for unfolding. These studies demonstrate the potential for generative AI models to improve physics analyses by reducing systematic uncertainties in detector corrections.

GAN-based unfolding methods face challenges in learning event correlations while handling detector acceptance effects. A recent approach [8] addresses these issues by incorporating a GAN with a pretrained classifier and a physics-informed loss function that penalizes the generator for deviations from true distributions. This improves topology classification and enhances unfolding accuracy, particularly in regions with limited detector acceptance.

4.2. *NF-based unfolding*

Beyond GANs, NFs have emerged as powerful tools for unfolding. Conditional invertible neural networks (cINNs) introduced the iterative conditional invertible neural network (IcINN) in [13] for single-event unfolding. Unlike traditional histogram-based unfolding, IcINNs operate at the event level, preserving detailed particle information while improving the handling of multidimensional correlations. Furthermore, their iterative nature reduces dependence on prior distributions used in Monte Carlo simulations, increasing adaptability across physics scenarios.

4.3. *DM-based unfolding*

DMs have gained traction as universal unfolding frameworks, mainly applying to denoise the detector smearing effects. DMs were first applied to multidimensional phase-space generation in 2023 [34], demonstrating improved generative modeling but facing challenges such as sparsity and discretization artifacts. Expanding on this, score-based diffusion models for point clouds [12] encoded particle-level information in a way that accommodates variable multiplicities and maintains correlations across observables. More recently, conditional denoising diffusion probabilistic models (cDDPMs) [55] have reframed unfolding as a denoising problem, treating detector-level data as a noisy version of the underlying truth. By incorporating distributional moments as conditioning information, cDDPMs generalize across multiple physics processes while preserving high-dimensional correlations. Schrödinger bridge-based unfolding (SBUnfold) [35] combines diffusion models with optimal transport to map detector-level events to truth-level events without explicit density estimation, improving performance in sparse phase spaces and enhancing robustness. OmniFold leverages iterative reweighting to simultaneously unfold all observables in high-dimensional, unbinned spaces [11]. DMs have also been explored for fast, high-fidelity data generation [60], including point-cloud diffusion to model hadronic jets [50].

A practical advantage of DMs is that they generate samples from the (truth-level) posterior distribution rather than a single deterministic output. This makes them well-suited to capturing the intrinsic randomness of the detection process and to propagating statistical fluctuations through the reverse chain. At the same time, the treatment of systematic uncertainties remains analysis-dependent. In practice, the same tools used in other unbinned unfolding approaches, such as bootstrapping, closure tests, alternative Monte Carlo samples, are still needed to validate the unfolding procedure and estimate uncertainties [11,55]. In this sense, DMs can provide a probabilistic description of the unfolding map, but they do not remove the need for a separate uncertainty assessment.

Another important issue is the potential mismatch between the detector simulation used in training and the true detector response in data. Because diffusion models learn the inverse map from simulated examples, their performance is only as reliable as the modeling of efficiencies, resolutions, acceptances, and correlations encoded in the training sample. For this reason, diffusion-based unfolding should be accompanied by dedicated validation studies, including closure tests, stress tests under detector variations, and comparisons with alternative event generators. This is particularly important in sparse regions or near sharp kinematic boundaries.

More generally, diffusion models are best viewed as a probabilistic unfolding engine rather than a substitute for the full experimental validation chain. Their stochastic sampling naturally encodes event-level ambiguity, and in this sense, they are particularly attractive for multidimensional observables

with strong correlations or variable multiplicities. However, a reliable physics result still requires the usual systematic studies, uncertainty propagation, and sanity checks needed to establish coverage and robustness.

Overall, DMs offer a promising framework for unbinned, multidimensional unfolding, especially when correlations and variable multiplicities are important, keeping present that their use to unfold real measurements requires proper standard validation and uncertainty studies.

4.4. *Physics-informed and moment-based unfolding*

Parallel to these advances, physics-informed neural networks (PINNs) have introduced a paradigm shift by embedding conservation laws and symmetries into the unfolding process. Although not inherently generative, PINNs enhance unfolding by constraining solutions to physically meaningful distributions. Originally introduced in 2017 [58], PINNs became increasingly utilized in particle physics by 2024 [63], improving the precision and interpretability of generative unfolding models. Furthermore, moment unfolding techniques [33] directly unfolded distribution moments, avoiding binning artifacts and improving jet substructure measurements. Despite these advances, extending moment unfolding to full-event reconstruction and integrating uncertainty quantification remain active areas of research.

As a high-level guide, GAN-based methods are often attractive when fast sampling and high-fidelity detector emulation are the primary goals, provided that training stability is carefully monitored. VAEs are useful when compact latent representations and controllable reconstruction are important. NFs are particularly valuable when likelihood evaluation, invertibility, or event-wise mappings are needed. DMs are well-suited for high-dimensional denoising and broad phase-space coverage, although their iterative sampling can be computationally expensive unless accelerated by distillation or related techniques. In practice, hybrid approaches such as latent diffusion, flow-matching, and VAE-GAN combinations increasingly blur the boundaries between these families by combining their complementary strengths.

5. Discussion

5.1. *Maximizing phase-space utilization via acceptance unfolding*

Unfolding data requires careful consideration of how different detector effects can introduce biases. Removing detector smearing within the accessible phase space can improve measurement resolution, and expanding the measured space enhances understanding of the underlying physics. Increased phase space coverage improves measured quantities by reducing extrapolation uncertainties and enabling more accurate descriptions of complex, multidimensional correlations [25]. This broader coverage is especially crucial in new physics searches, where subtle deviations in high-dimensional observables may signal beyond-the-standard-model effects [19]. One promising approach involves linking different phase space regions covered by various experiments. However, this method faces significant challenges, as each experiment introduces unique biases and systematic uncertainties that must be carefully accounted for. Consequently, results are often limited to data from a single experimental setup. Successful extraction of relevant physical variables from multiple experiments, coupled with accurate unfolding of their respective detector effects, could yield a unified dataset

spanning the largest possible portion of the available phase space. This would, in turn, enhance measurement precision and reliability.

5.2. Latent space

Another unique feature of AI-supported data analysis is its ability to investigate latent space, seeking a correspondence between the neural network's weight distribution and the underlying physics mechanisms. A well-known approach for this is the t-distributed stochastic neighbor embedding (t-SNE) method [62], which takes a set of points in a high-dimensional space and finds a faithful representation of those points in a lower-dimensional space, typically a 2D plane.

Preliminary studies on two-pion photoproduction on protons in CLAS [7] indicate that the latent space of a GAN is selectively sensitive to different production mechanisms, as shown in Figure 3. It is well-known that this reaction can proceed through different intermediate states that excite mesons or baryons. After training the neural network with pseudodata generated from several reaction models, the response to a subset containing only one mechanism clustered in specific regions of the t-SNE space. This suggests that particular areas of the latent space become activated for different production mechanisms. In principle, by reversing the procedure, it could be possible to disentangle the different production mechanisms (e.g., intermediate meson and baryon resonances) by isolating their corresponding regions in latent space. Despite these promising results, further studies are needed. The structure of t-SNE diagrams depends heavily on hyperparameter choices, such as perplexity, learning rate, the number of iteration steps, and even variations across different runs. A thorough analysis is currently underway to determine whether the observed t-SNE patterns truly reflect different physics mechanisms rather than being artifacts of the nonlinear t-SNE visualization.

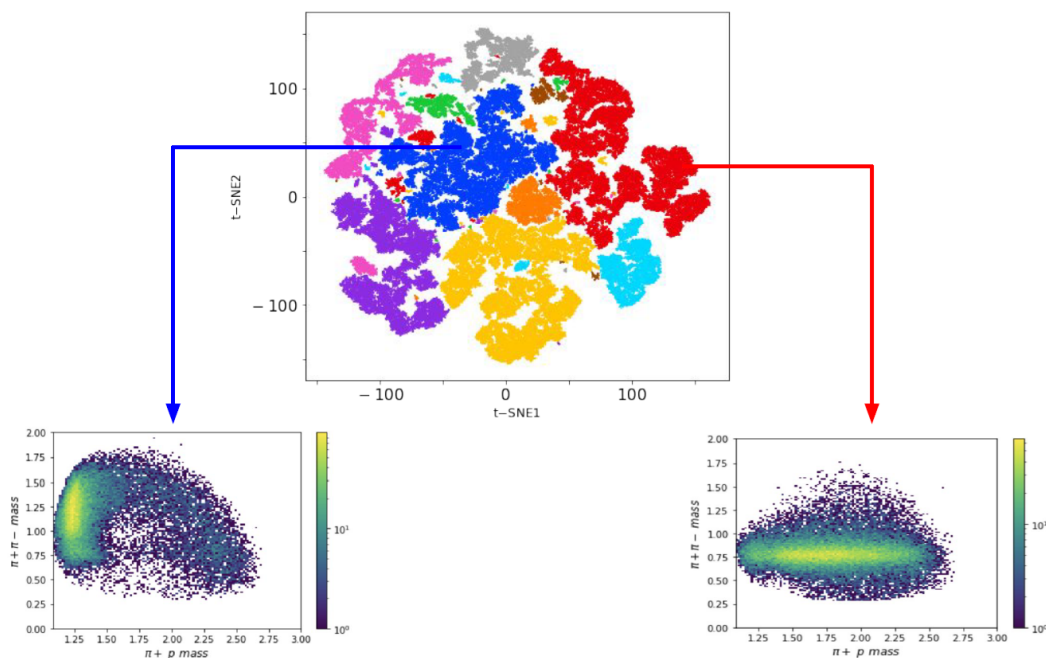


Figure 3. Latent-space clusters from a CLAS two-pion photoproduction on protons sample; clusters correspond to different resonances.

5.3. Background removal

A closely related problem to unfolding detector effects is removing the background, which is also present in an event sample. This correlates with detector issues such as resolution and acceptance in that the larger these issues are the greater the background contribution is likely to be. Removing this background is also required to extract the true vertex level event distributions and hence physics parameters of interest. One approach with generative AI is to generate distributions for both signal and background simultaneously, only using the signal distribution for physics extraction. This is likely to prove difficult as will require modeling the background which often comes from a multitude of sources.

Current methods often subtract isolated background event samples via negative event weighting, such as “sideband subtractions” or the sPlot technique [57]. In principle, using such weights in generative model training could present a solution to this problem. However, loss functions often do not deal correctly with negative sample weights, leading to alternative, or intermediate, approaches. One method of transforming weights to positive definite was suggested by [51]. Recently, Glazier and Tyson [38] used a similar method to transform sPlot weights to probabilities to remove backgrounds in ML training. This essentially requires performing a probabilistic classification based on two datasets, one with signal weights applied, the other containing all events. The output from this can then be converted to the probability that the event is from the signal class and used as a weight for training events.

5.4. Uncertainty quantification (UQ)

UQ is essential for ensuring the reliability of generative models, particularly in the context of particle physics simulations. It helps assess the accuracy and trustworthiness of model predictions, which is crucial for decision-making in experiments.

Several methods have been proposed to assess uncertainty in generative models. For instance, Anderlini et al. [10] introduced techniques to quantify model uncertainty both within and beyond the training phase space, using data-driven calibration methods. This approach has been applied to Large Hadron Collider beauty ring-imaging Cherenkov detector (LHCb RICH) fast simulations, improving the model’s reliability in particle detection. In [47], ensemble methods were explored for classification tasks, demonstrating that they can provide well-calibrated classification probabilities. This technique enhances the robustness of particle classification, which is crucial for high-precision physics experiments. Finally, Fanelli and Giroux [36] presented a physics-informed Bayesian neural network, which uses flow-approximated posteriors for detailed event-level uncertainty quantification. This model is particularly effective for extracting kinematic variables and provides crucial uncertainty estimates, aiding in tasks such as event filtering and anomaly detection.

These studies highlight different approaches to managing uncertainty across generative models, with applications ranging from fast simulations to detailed event-level analysis, offering significant improvements in model reliability for particle physics.

5.5. Extracting physics insights

From an ML perspective, extracting scientific knowledge from experimental data requires more than the direct application of off-the-shelf algorithms. Many problems in particle physics can be naturally formulated as ill-posed, high-dimensional inverse inference tasks, where the objective is

to recover latent generative factors from incomplete, noisy, and biased observations. In this setting, incorporating domain knowledge as inductive bias is critical for enabling meaningful inference beyond purely correlational predictions. Physics-informed ML addresses this challenge by embedding known constraints, such as conservation laws, symmetries, and structural relationships, into ML model architectures, training objectives, or probabilistic priors, thereby guiding learning toward physically consistent solutions [1].

In this framework, generative models have emerged as a powerful class of methods for solving complex inverse problems at scale. By learning expressive latent representations and flexible conditional distributions, physics-informed generative models enable data-driven inference of hidden variables characterized by high dimensionality, strong correlations, and limited supervision. These models are well-suited for reconstructing latent processes from experimental observations, providing a principled approach to uncertainty-aware inference without relying on restrictive analytical models.

In hadron spectroscopy, existing ML applications have largely focused on supervised classification tasks, such as labeling or discriminating candidate hadronic states [52, 61]. In contrast, generative models have so far been primarily applied at the level of observable probability distributions (e.g., cross sections), rather than at the level of latent representations that encode quantum interference effects. Recovering such structured latent variables poses a substantially more challenging inference problem. Recent exploratory works have begun to address this gap through physics-constrained GANs that incorporate domain constraints. Focusing on a simplified $2 \rightarrow 2$ elastic scattering setting, these models enforce unitarity constraints, effectively regularizing the inverse problem and enabling stable reconstruction of latent amplitudes from observed cross sections. Generative AI-based surrogates simulating physics event sampling [6] on top of detector unfolding have also been developed to obtain parameterization of the quantum correlation functions.

6. Conclusions and outlook

In this work, we have explored the impact of generative models, such as GANs, VAEs, NFs, and DMs, on detector simulations. These approaches offer significant advantages in improving the speed and accuracy of detector simulations while preserving the underlying physics. Leveraging these advanced generative models, we can achieve more efficient and scalable detector simulations, reducing reliance on computationally expensive Monte Carlo methods.

Three main lessons emerge from this survey. First, detector-effect unfolding should be viewed as an uncertainty-aware generative inverse problem rather than only as a post-processing correction to measured histograms. Second, no single generative architecture dominates all use cases: GANs often provide fast high-fidelity sampling, NFs offer invertible and likelihood-aware mappings, DMs provide robust denoising and phase-space coverage, and VAEs remain useful for compact latent representations. In practice, hybrid approaches can also combine the complementary strengths of these methods. Third, the next major progress will likely come from physically constrained and uncertainty-calibrated models that preserve event-level correlations while remaining sufficiently interpretable, supported by rigorous closure tests, to be trusted in precision experimental analyses.

Our analysis of unfolding and folding techniques highlights their potential for refining data interpretation and enhancing physics analyses. Specifically, we have demonstrated how generative models can be used to mitigate detector effects, enabling a more direct comparison between

experimental data and theoretical predictions. The implementation of acceptance unfolding strategies has further allowed for maximizing phase-space utilization, ensuring that all relevant information is retained for downstream analyses.

Despite these advancements, several challenges remain. Ensuring robust UQ is critical to maintaining the credibility of ML-driven approaches in physics. Current methods, such as Bayesian deep learning and adversarial training, need further refinement to provide reliable confidence estimates. Additionally, improving model interpretability remains a key area of focus, as understanding latent representations and their correlation with physical features is essential for trust and adoption in experimental workflows.

Integrating generative models into real-world experimental setups presents additional challenges, including addressing computational constraints, optimizing model inference times, and standardizing validation techniques across different detectors and experimental conditions. Future work should focus on hybrid approaches that combine the strengths of different generative architectures to further enhance detector effects unfolding performance and robustness.

Ultimately, the generative models hold great promise for transforming particle and nuclear physics by improving simulation accuracy, accelerating data analysis, uncovering new physics insights, and advancing theories. Continued AI research and interdisciplinary cooperation will be crucial in realizing their full potential and shaping the future of scientific discovery.

Acknowledgments

This work is partially supported by the U.S. Department of Energy, Office of Science, Office of Nuclear Physics, Office of Advanced Scientific Computing Research through the Scientific Discovery through Advanced Computing (SciDAC) program, under contracts DE-AC02-06CH11357, DE-AC05-06OR23177, and DE-SC0023472, for the award *Femtосcale Imaging of Nuclei using Exascale Platforms*; the EXCLAIM collaboration under the DOE grants DE-SC0016286 and DE-SC0024644; the Center for Nuclear Femtography (CNF), administrated by the Southeastern Universities Research Association under an appropriation from the Commonwealth of Virginia under contract No. C2024-FEMT-011-02; the ODU Data Science Seed Fundings 2023; and the U.K. Science and Technology Facilities Council under grant ST/V00106X/1. We acknowledge the support of the GeV-AI project (CUP I57G21000110007), in the context of the ICSC Spoke 2 Open Calls, project funded by European Union – NextGenerationEU – and National Recovery and Resilience Plan (NRRP) – Mission 4 Component 2.

Use of AI tools declaration

The authors declare that generative Artificial Intelligence (AI) tools were used solely to improve the language and clarity of the manuscript. No AI tools were used for content generation, data analysis, or interpretation. All scientific work, tables, and findings were created and validated exclusively by the authors.

Conflict of interest

Yaohang Li is an editorial board member for Applied Computing and Intelligence and was not involved in the editorial review and the decision to publish this article.

References

1. G. Aarts, K. Fukushima, T. Hatsuda, A. Ipp, S. Shi, L. Wang, et al., Physics-driven learning for inverse problems in quantum chromodynamics, *Nat. Rev. Phys.*, **7** (2025), 154–163. <https://doi.org/10.1038/s42254-024-00798-x>
2. A. Abhishek, E. Drechsler, W. Fedorko, B. Stelzer, CaloDVAE: discrete variational autoencoders for fast calorimeter shower simulation, *Proceedings of the Fourth Workshop on Machine Learning and the Physical Sciences*, 2021, 1–11.
3. Y. Alanazi, N. Sato, P. Ambrozewicz, A. Hiller Blin, W. Melnitchouk, M. Battaglieri, et al., A survey of machine learning-based physics event generation, *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI-21)*, 2021, 4286–4293. <https://doi.org/10.24963/ijcai.2021/588>
4. Y. Alanazi, P. Ambrozewicz, M. Battaglieri, A. Hiller Blin, M. Kuchera, Y. Li et al., Machine learning-based event generator for electron-proton scattering, *Phys. Rev. D*, **106** (2022), 096002. <https://doi.org/10.1103/PhysRevD.106.096002>
5. Y. Alanazi, N. Sato, T. Liu, W. Melnitchouk, P. Ambrozewicz, F. Hauenstein, et al., Simulation of electron-proton scattering events by a feature-augmented and transformed generative adversarial network (FAT-GAN), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)*, 2021, 2126–2132. <https://doi.org/10.24963/ijcai.2021/293>
6. T. Alghamdi, J. Xu, N. Ramachandra, N. Sato, Y. Li, Towards an event-level analysis in hadronic physics using generative ai-based surrogates, *Proceedings of IEEE 37th International Conference on Tools with Artificial Intelligence (ICTAI)*, 2025, 445–452. <https://doi.org/10.1109/ICTAI66417.2025.00067>
7. T. Alghamdi, Y. Alanazi, M. Battaglieri, Ł. Bibrzycki, A. V. Golda, A. N. Hiller Blin, et al., Toward a generative modeling analysis of CLAS exclusive 2π photoproduction, *Phys. Rev. D*, **108** (2023), 094030. <https://doi.org/10.1103/PhysRevD.108.094030>
8. T. Alghamdi, T. Vittorini, M. Spreafico, M. Battaglieri, N. Sato, Y. Li, Unfolding particle detector acceptance in high energy physics with generative AI, *Proceedings of IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI)*, 2024, 962–968. <https://doi.org/10.1109/ICTAI62512.2024.00138>
9. O. Amram, K. Pedro, Denoising diffusion models with geometry adaptation for high fidelity calorimeter simulation, *Phys. Rev. D*, **108** (2023), 072014. <https://doi.org/10.1103/PhysRevD.108.072014>
10. L. Anderlini, C. Chimpoesh, N. Kazeev, A. Shishigina, Generative models uncertainty estimation, *J. Phys.: Conf. Ser.*, **2438** (2023). 012088. <https://doi.org/10.1088/1742-6596/2438/1/012088>

11. A. Andreassen, P. Komiske, E. Metodiev, B. Nachman, J. Thaler, Omnifold: a method to simultaneously unfold all observables, *Phys. Rev. Lett.*, **124** (2020), 182001. <https://doi.org/10.1103/PhysRevLett.124.182001>
12. J. Araz, V. Mikuni, F. Ringer, N. Sato, F. Acosta, R. Whitehill, Point cloud-based diffusion models for the Electron-Ion Collider, *Phys. Lett. B*, **868** (2025), 139694. <https://doi.org/10.1016/j.physletb.2025.139694>
13. M. Backes, A. Butter, M. Dunford, B. Malaescu, Event-by-event comparison between machine-learning- and transfer-matrix-based unfolding methods, *Eur. Phys. J. C*, **84** (2024), 770. <https://doi.org/10.1140/epjc/s10052-024-13136-3>
14. H. Baran, *Deep generative models for ultra-high granularity particle physics detector simulation: a voyage from emulation to extrapolation*, Ph. D Thesis, Ludwig-Maximilians-Universität München, 2023. <https://doi.org/10.5282/edoc.34137>
15. M. Bellagente, A. Butter, G. Kasieczka, T. Plehn, R. Winterhalder, How to GAN away detector effects, *SciPost Phys.*, **8** (2020), 070. <https://doi.org/10.21468/SciPostPhys.8.4.070>
16. S. Biedron, L. Brouwer, D. L. Bruhwiler, N. M. Cook, A. L. Edelen, D. Filippetto, et al., Snowmass21 accelerator modeling community white paper, arXiv: 2203.08335. <https://doi.org/10.48550/arXiv.2203.08335>
17. V. Blobel, An unfolding method for high-energy physics experiments, *Proceedings of the Conference on Advanced Statistical Techniques in Particle Physics*, 2002, 258–267.
18. A. Boehnlein, M. Diefenthaler, N. Sato, M. Schram, V. Ziegler, C. Fanelli, et al., Colloquium: machine learning in nuclear physics, *Rev. Mod. Phys.*, **94** (2022), 031003. <https://doi.org/10.1103/revmodphys.94.031003>
19. J. Brehmer, G. Louppe, J. Pavez, K. Cranmer, Mining gold from implicit models to improve likelihood-free inference, *Proc. Natl. Acad. Sci. U.S.A.*, **117** (2020) 5242–5249. <https://doi.org/10.1073/pnas.1915980117>
20. E. Buhmann, S. Diefenbacher, E. Eren, F. Gaede, G. Kasieczka, A. Korol, et al., Caloclouds: fast geometry-independent highly-granular calorimeter simulation, *JINST*, **18** (2023), P11025. <https://doi.org/10.1088/1748-0221/18/11/P11025>
21. E. Buhmann, S. Diefenbacher, E. Eren, F. Gaede, G. Kasieczka, A. Korol, et al., Decoding photons: physics in the latent space of a BIB-AE generative network, *EPJ Web Conf.*, **251** (2021), 03003. <https://doi.org/10.1051/epjconf/202125103003>
22. E. Buhmann, F. Gaede, G. Kasieczka, A. Korol, W. Korcari, K. Krüger, et al., CaloClouds II: ultra-fast geometry-independent highly-granular calorimeter simulation, *JINST*, **19** (2024), P04020. <https://doi.org/10.1088/1748-0221/19/04/P04020>
23. A. Butter, T. Plehn, Generative networks for LHC events, In: *Artificial intelligence for high energy physics*, Singapore: World Scientific Publishing, 2020, 191–240. https://doi.org/10.1142/9789811234033_0007
24. G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, et al., Machine learning and the physical sciences, *Rev. Mod. Phys.*, **91** (2019), 045002. <https://doi.org/10.1103/RevModPhys.91.045002>

25. J. Chan, B. Nachman, Unbinned profiled unfolding, *Phys. Rev. D*, **108** (2023), 016002. <https://doi.org/10.1103/PhysRevD.108.016002>
26. S. V. Chekanov, HepSim: a repository with predictions for high-energy physics experiments, *Adv. High Energy Phys.*, **2015** (2015), 136093. <https://doi.org/10.1155/2015/136093>
27. G. Cowan, *Statistical data analysis*, Oxford: Oxford Academic, 1998. <https://doi.org/10.1093/oso/9780198501565.001.0001>
28. J. Cresswell, B. Ross, G. Loaiza-Ganem, H. Reyes-Gonzalez, M. Letizia, A. Caterini, Caloman: fast generation of calorimeter showers with density estimation on learned manifolds, *Proceedings of the Machine Learning and the Physical Sciences Workshop*, 2022, 1–8.
29. K. Datta, D. Kar, D. Roy, Unfolding with generative adversarial networks, arXiv: 1806.00433. <https://doi.org/10.48550/arXiv.1806.00433>
30. L. de Oliveira, M. Paganini, B. Nachman, Learning particle physics by example: location-aware generative adversarial networks for physics synthesis, *Comput. Softw. Big Sci.*, **1** (2017), 4. <https://doi.org/10.1007/s41781-017-0004-6>
31. M. Debbagh, Learning structured output representations from attributes using deep conditional generative models, arXiv: 2305.00980. <https://doi.org/10.48550/arXiv.2305.00980>
32. K. Deja, J. Dubiński, P. Nowak, S. Wenzel, P. Spurek, T. Trzcinski, End-to-end sinkhorn autoencoder with noise generator, *IEEE Access*, **9** (2020), 7211–7219. <https://doi.org/10.1109/ACCESS.2020.3048622>
33. K. Desai, B. Nachman, J. Thaler, Moment extraction using an unfolding protocol without binning, *Phys. Rev. D*, **110** (2024), 116013. <https://doi.org/10.1103/PhysRevD.110.116013>
34. P. Devlin, J. W. Qiu, F. Ringer, N. Sato, Diffusion model approach to simulating electron-proton scattering events, *Phys. Rev. D*, **110** (2024), 016030. <https://doi.org/10.1103/PhysRevD.110.016030>
35. S. Diefenbacher, G. Liu, V. Mikuni, B. Nachman, W. Nie, Improving generative model-based unfolding with schrödinger bridges, *Phys. Rev. D*, **109** (2024), 076011, <https://doi.org/10.1103/PhysRevD.109.076011>
36. C. Fanelli, J. Giroux, ELUQuant: event-level uncertainty quantification in deep inelastic scattering, *Mach. Learn.: Sci. Technol.*, **5** (2024), 015017. <https://doi.org/10.1088/2632-2153/ad2098>
37. M. F. Giannelli, R. Zhang, CaloShowerGAN, a generative adversarial network model for fast calorimeter shower simulation, *Eur. Phys. J. Plus*, **139** (2024), 597. <https://doi.org/10.1140/epjp/s13360-024-05397-4>
38. D. Glazier, R. Tyson, Converting sweights to probabilities with density ratios, *Comput. Phys. Commun.*, **318** (2026), 109890. <https://doi.org/10.1016/j.cpc.2025.109890>
39. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., Generative adversarial nets, *Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2014, 2672–2680.

40. A. Hariri, D. Dyachkova, S. Gleyzer, Graph generative models for fast detector simulations in high energy physics, *Proceedings of the Third Workshop on Machine Learning and the Physical Sciences*, 2021, 1–6.
41. J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, *Proceedings of the 34th Conference on Neural Information Processing Systems*, 2020, 1–12.
42. Y. Huang, Y. Ren, S. Yoo, J. Huang, Fast 2D bicephalous convolutional autoencoder for compressing 3D time projection chamber data, *Proceedings of the SC '23 Workshops of the International Conference on High Performance Computing, Network, Storage, and Analysis*, 2023, 298–305. <https://doi.org/10.1145/3624062.3625127>
43. K. Jaruskova, S. Vallecorsa, Ensemble models for calorimeter simulations, *J. Phys.: Conf. Ser.*, **2438** (2023), 012080. <https://doi.org/10.1088/1742-6596/2438/1/012080>
44. R. Kansal, J. Duarte, B. Orzari, T. Tomei, M. Pierini, M. Touranakou, et al., Graph generative adversarial networks for sparse data generation in high energy physics, *Proceedings of the Third Workshop on Machine Learning and the Physical Sciences*, 2021, 1–9.
45. D. P. Kingma, M. Welling, Auto-encoding variational bayes, *Proceedings of the International Conference on Learning Representations*, 2014, 1–14.
46. D. Kobylanskii, N. Soybelman, N. Kakati, E. Dreyer, B. Nachman, E. Gross, Advancing set-conditional set generation: diffusion models for fast simulation of reconstructed particles, *Phys. Rev. D*, **110** (2024), 092013. <https://doi.org/10.1103/PhysRevD.110.092013>
47. D. Koh, A. Mishra, K. Terao, Deep neural network uncertainty quantification for LArTPC reconstruction, *JINST*, **18** (2023), P12013. <https://doi.org/10.1088/1748-0221/18/12/P12013>
48. C. Krause, D. Shih, CaloFlow II: even faster and still accurate generation of calorimeter showers with normalizing flows, arXiv: 2110.11377. <https://doi.org/10.48550/arXiv.2110.11377>
49. C. Krause, D. Shih, Fast and accurate simulations of calorimeter showers with normalizing flows, *Phys. Rev. D*, **107** (2023), 113003. <https://doi.org/10.1103/PhysRevD.107.113003>
50. V. Mikuni, B. Nachman, M. Pettee, Fast point cloud generation with diffusion models in high energy physics, *Phys. Rev. D*, **108** (2023), 036025. <https://doi.org/10.1103/PhysRevD.108.036025>
51. B. Nachman, J. Thaler, Neural resampler for Monte Carlo reweighting with preserved uncertainties, *Phys. Rev. D*, **102** (2020), 076004. <https://doi.org/10.1103/PhysRevD.102.076004>
52. L. Ng, Ł. Bibrzycki, J. Nys, C. Fernández-Ramírez, A. Pilloni, V. Mathieu, et al., Deep learning exotic hadrons, *Phys. Rev. D*, **105** (2022), L091501. <https://doi.org/10.1103/PhysRevD.105.L091501>
53. M. Paganini, L. de Oliveira, B. Nachman, CaloGAN: simulating 3D high energy particle showers in multilayer electromagnetic calorimeters with generative adversarial networks, *Phys. Rev. D*, **97** (2018), 014021. <https://doi.org/10.1103/PhysRevD.97.014021>
54. I. Pang, D. Shih, J. A. Raine, Calorimeter shower superresolution, *Phys. Rev. D*, **109** (2024), 092009. <https://doi.org/10.1103/PhysRevD.109.092009>

55. C. Pazos, S. Aeron, P. Beauchemin, V. Croft, Z. Huan, M. Klassen, et al., Towards universal unfolding of detector effects in high-energy physics using denoising diffusion probabilistic models, arXiv: 2406.01507. <https://doi.org/10.48550/arXiv.2406.01507>
56. A. Peisert, F. Sauli, *Drift and diffusion of electrons in gases: a compilation (with an introduction to the use of computing programs)*, Geneva: CERN, 1984. <https://doi.org/10.5170/CERN-1984-008>
57. M. Pivk, F. R. Le Diberder, SPlot: a statistical tool to unfold data distributions, *Nucl. Instrum. Meth. A*, **555** (2005), 356–369. <https://doi.org/10.1016/j.nima.2005.08.106>
58. M. Raissi, P. Perdikaris, G. Em Karniadakis, Physics informed deep learning (part I): data-driven solutions of nonlinear partial differential equations, arXiv: 1711.10561. <https://doi.org/10.48550/arXiv.1711.10561>
59. D. J. Rezende, S. Mohamed, Variational inference with normalizing flows, *Proceedings of the 32nd International Conference on Machine Learning*, 2015, 1530–1538.
60. T. Salimans, J. Ho, Progressive distillation for fast sampling of diffusion models, arXiv: 2202.00512. <https://doi.org/10.48550/arXiv.2202.00512>
61. D. L. B. Sombillo, Y. Ikeda, T. Sato, A. Hosaka, Classifying the pole of an amplitude using a deep neural network, *Phys. Rev. D*, **102** (2020), 016024. <https://doi.org/10.1103/PhysRevD.102.016024>
62. L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, *J. Mach. Learn. Res.*, **9** (2008), 2579–2605.
63. V. Vatellis, Advancing physics data analysis through machine learning and physics-informed neural networks, arXiv: 2410.14760. <https://doi.org/10.48550/arXiv.2410.14760>
64. S. Zhao, J. Song, S. Ermon, InfoVAE: balancing learning and inference in variational autoencoders, *Proceedings of the AAAI Conference on Artificial Intelligence*, **33** (2019), 5885–5892. <https://doi.org/10.1609/aaai.v33i01.33015885>



AIMS Press

© 2026 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>)