PROCEEDINGS OF SCIENCE

Towards Universal Unfolding using Denoising Diffusion

Camila Pazos,[a,*] **Shuchin Aeron,**[b,d] **Pierre-Hugues Beauchemin,**[a,d] **Vincent Croft,**[c] **Martin Klassen**[a] **and Taritree Wongjirad**[a,d]

[a] *Department of Physics and Astronomy, Tufts University, Medford, Massachusetts*
[b] *Department of Electrical and Computer Engineering, Tufts University, Medford, Massachusetts*
[c] *Leiden Institute for Advanced Computer Science LIACS, Leiden University, The Netherlands*
[d] *The NSF AI Institute for Artificial Intelligence and Fundamental Interactions*
E-mail: camila.pazos@tufts.edu, shuchin@eecs.tufts.edu, hugo.beauchemin@tufts.edu, vincent.croft@cern.ch, martin.klassen@tufts.edu, taritree.wongjirad@tufts.edu

Correcting for detector effects in experimental data, particularly through unfolding, is critical for enabling precision measurements in high-energy physics. However, traditional unfolding methods face challenges in scalability, flexibility, and dependence on simulations. We introduce a novel approach to multidimensional object-wise unfolding using conditional Denoising Diffusion Probabilistic Models (cDDPM). Our method utilizes the cDDPM for a non-iterative, flexible posterior sampling approach, incorporating distribution moments as conditioning information, which exhibits a strong inductive bias that allows it to generalize to unseen physics processes without explicitly assuming the underlying distribution. Our results highlight the potential of this method as a step towards a "universal" unfolding tool that reduces dependence on truth-level assumptions, while enabling the unfolding of a wide range of measured distributions with enhanced flexibility and accuracy.

12th Edition of the Large Hadron Collider Physics Conference (LHCP2024)
3-7 June 2024
Boston, USA

[*] Speaker

PoS(LHCP2024)234

1. Introduction

Experimental data in high-energy physics (HEP) presents a distorted picture of the true physics processes due to detector effects. Unfolding is an inverse problem solved through statistical inference that aims to correct the detector distortions of the observed data to recover the true distribution of particle properties. This process is essential for the validation of theories, new discoveries, precision measurements, and comparison of experimental results between different experiments.

A standard approach to unfolding [1] begins with a predicted particle distribution $f_{\text{true}}(x)$ that characterizes the underlying physics process of interest, and a detailed detector simulation that describes how detector effects distort the particle property distributions. These distortions affect the kinematic quantities of particles incident to the detector, resulting in an altered particle distribution $f_{\text{det}}(y)$. This relationship can be written as a Fredholm integral equation of the first kind,

$$f_{\text{det}}(y) = \int dx\, P(y|x)\, f_{\text{true}}(x) \tag{1}$$

where $P(y|x)$ is the conditional probability distribution describing the detector effects. Unfolding requires the inverse process $P(x|y)$, which can be expressed with Bayes' theorem as

$$P(x|y) = \frac{P(y|x) f_{\text{true}}(x)}{f_{\text{det}}(y)} \,. \tag{2}$$

In this context, a detector dataset can be unfolded by sampling from the posterior $P(x|y)$ to recover the distribution $f_{\text{true}}(x)$. The detector effects $P(y|x)$ are assumed to be the same for any physics process, and we can see that the posterior $P(x|y)$ depends on the prior distribution $f_{\text{true}}(x)$. Although we can sample from $f_{\text{true}}(x)$ through the use of particle generators, there is no guarantee that any particular assumed $f_{\text{true}}(x)$ accurately represents the underlying physics of the specific data we want to unfold. Consequently, unfolding results can be significantly influenced by the assumed underlying distribution, potentially introducing bias or limiting the method's ability to detect unexpected phenomena. This reveals one of the main challenges in developing a *universal* unfolder, which aims to remove detector effects from any set of measured data agnostic of the process of interest, ideally with no bias towards any prior distribution. Additionally, traditional unfolding methods, based on the linearization of the problem, face limitations such as requiring binned histograms and an inability to unfold multiple observables simultaneously.

Related Work: Various machine learning approaches have emerged in recent years to address these challenges. These include re-weighting methods like OmniFold [2] [3], as well as several generative approaches. Among the generative techniques are those using Generative Adversarial Networks (GANs) [4] [5], conditional invertible neural networks [6] [7], and latent variational diffusion models [8] [9]. Additionally, distribution mapping techniques have been developed, such as Schrödinger bridges [10] and direct diffusion models [11]. For a comprehensive overview of these methods, see the recent survey by [12]. Each new method has made further strides in unfolding and shown the advantages in machine learning based approaches compared to traditional techniques.

Our Contribution: In this work, we introduce a novel approach based on conditional Denoising Diffusion Probabilistic Models (cDDPM) to unfold detector effects in HEP data. We demonstrate that a single cDDPM, trained on particle data from a diverse range of physics distributions and incorporating the statistical moments of these distributions, can be used as a "generalized" unfolder to perform multidimensional object-wise unfolding for a variety of physics processes without requiring an explicit assumption about the underlying distribution that could bias the results. Figure 1 shows that this generalized unfolder outperforms the standard "dedicated" unfolding approach for unknown physics processes (details in Section 3). This flexibility is crucial for new physics searches and studying processes not accurately modeled by current theories. Moreover, our approach eliminates the need for process-specific model training, enabling the unfolding of a million data points in approximately 3 minutes.

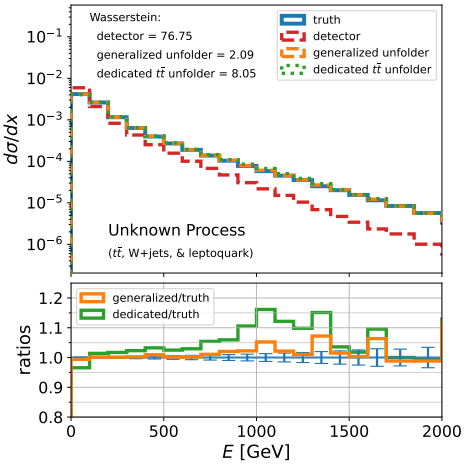

Figure 1: Unfolding results for an "unknown" process, comparing our generalized cDDPM unfolder against the standard, dedicated unfolding approach.

2. Methods

2.1 Our Unfolding Approach

Although we cannot achieve an ideal universal unfolder, we seek an approach that will enhance the inductive bias of the unfolding method to improve generalization to cover various posteriors pertaining to different physics data distributions. We can see that the posteriors for two different physics processes i and j are related by a ratio of the probability density functions of each process,

$$\frac{P_i(x|y)}{P_j(x|y)} = \frac{f^i_{\text{true}}(x)\, f^j_{\text{det}}(y)}{f^i_{\text{det}}(y)\, f^j_{\text{true}}(x)} \,. \tag{3}$$

Assuming we can learn the posterior for a given physics process, we could extrapolate to unseen posteriors if the priors $f_{\text{true}}(x)$ and detector distributions $f_{\text{det}}(y)$ can be approximated or written in a closed form. Although these functions have no analytical form, we can approximate key features using the first moments of these distributions. By making use of these moments as conditionals, we can have a more flexible unfolder that is not strictly tied to a selected prior distribution, and enables it to interpolate and extrapolate to unseen posteriors. Consequently, this unfolding tool gains the ability to handle a wider range of physics processes and enhances the generalization capabilities, making it a more versatile tool for unfolding in various high energy physics applications.

2.2 Denoising Diffusion Probabilistic Models

Our proposed unfolding approach calls for a flexible generative model, and denoising diffusion probabilistic models (DDPMs) [13] lend themselves naturally to this task. DDPMs learn via a reversible generative process which we can condition directly on the simulated detector data values

and on the moments of the distribution $f_{\text{det}}(y)$, providing a natural way to sample from $P(x|y)$ for unfolding. A DDPM comprises two parts: a fixed forward process that gradually adds Gaussian noise to data samples, and a learned reverse process that denoises the data.

We implement a conditional DDPM (cDDPM) with direct conditioning, where sampling is done according to the learned conditional distribution. The cDDPM allows us to sample from $P(x|y)$ without explicitly evaluating the prior distribution over the data space, making it a natural choice for applications like unfolding where the prior is unknown or difficult to model.

2.3 Unfolding with cDDPMs

Our study focuses on QCD jets. Using the PYTHIA event generator [14], we generate jet datasets for various physics processes ($t\bar{t}$, W+jets, Z+jets, dijet, and leptoquark) under different settings. These "truth-level" jets are then passed through a detector simulation framework to produce "detector-level" jets, mimicking particle interactions within a detector.

Part 1: Dedicated Unfolder We first consider how to setup a *dedicated* cDDPM unfolder (without use of the distributional moments) that can achieve multidimensional object-wise unfolding for a single physics process. The jet kinematic information is defined with a vector that includes the transverse momentum (p_T), pseudorapidity (η), azimuthal angle (ϕ), and 4-momentum vector (E, p_x, p_y, p_z). These jet vectors are defined both at truth-level as $\vec{x}$ and detector-level as $\vec{y}$. A cDDPM can be trained with data pairs $(\vec{x}, \vec{y})$ as input to learn the posterior distribution $P(\vec{x}|\vec{y})$. To unfold, we give the detector data $\vec{y}$ as input and the cDDPM acts as a posterior sampler of $P(\vec{x}|\vec{y})$.

Part 2: Generalized Unfolder We aim to enhance the inductive bias through use of the distributional moments to attain a *generalized* cDDPM unfolder that encompasses a broader range of posteriors, enabling the unfolding of data from diverse physics processes. To achieve this, we expand our training dataset to include jets from multiple different physics simulations. For each simulation, we compute the first 6 moments of the p_T distribution and append them to the corresponding jet vectors. We use the first 6 moments as our tests showed this to be the minimum number yielding optimal results. In slight abuse of notation, we now denote these augmented jet vectors (including distribution moments) as $\vec{x}$ at truth-level and $\vec{y}$ at detector-level. By training with these diverse data pairs $(\vec{x}, \vec{y})$, we enable the cDDPM to represent multiple posteriors corresponding to the distributions in the expanded training dataset, distinguishable through the added distributional information provided by the moments.

3. Results and Discussion

To evaluate our unfolding approach, we employ the Wasserstein distance [15] as a metric, comparing the discrepancy between truth and unfolded values against that of truth and detector values. Figure 1 showcases results from an "unknown" process dataset, created by combining jets from the $t\bar{t}$, W+jets, and leptoquark test datasets. Our generalized unfolder demonstrates superior performance when unfolding this unknown process compared to a dedicated unfolder assuming a similar, yet incorrect, underlying $t\bar{t}$ process. While the generalized unfolder's advantage is expected for unknown processes, we also aim for comparable performance to dedicated unfolders on known processes. To validate our framework's effectiveness we compare both unfolders across various test

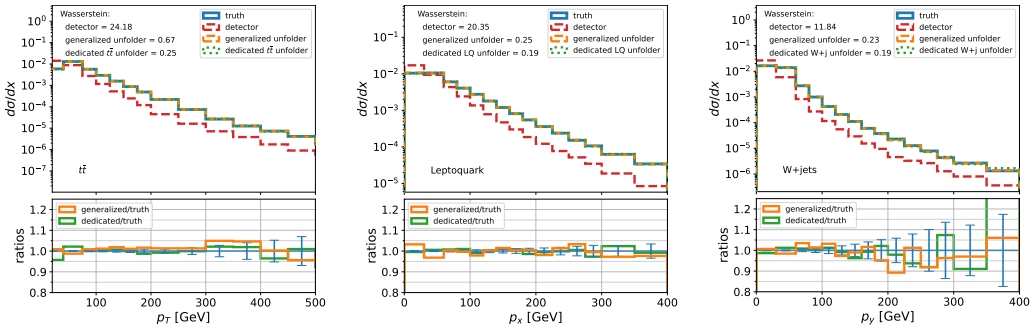

Figure 2: Unfolding results of jet vector components across diverse physics processes. We compare our generalized cDDPM unfolder (orange) against process-specific dedicated unfolders (green). Deviations from one are within the expected uncertainty budget typical of experimental measurements of these distributions.

datasets, and Table 1 presents the resulting multidimensional Wasserstein distances to their true distributions.

Figure 2 illustrates unfolding results for various jet observables across different physics processes, showcasing the generalized unfolder's versatility. In Figure 3, the model's efficacy is further demonstrated with two tests: (1) reconstructing jet mass from unfolded results, indicating well-preserved correlations among jet vector components, and (2) reconstructing event-level observables from unfolded quantities, achieved by tracking event numbers through object-wise unfolding.

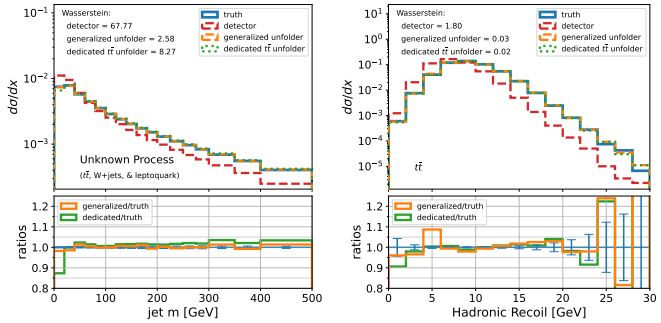

Figure 3: Reconstruction of jet mass and hadronic recoil (event-level observable) from unfolded data.

Process	Wasserstein Distances		
	Det.	Gen.	Ded.
Unknown	28.20	0.744	2.677
$t\bar{t}$	26.43	0.565	0.196
LQ	32.72	0.457	0.155
W+jets	31.15	0.304	0.353

Table 1: Comparison of Wasserstein distances for detector-level data and unfolded results using generalized and dedicated unfolders across different physics processes.

While this approach shows promise, we acknowledge key limitations. Addressing particles outside detector thresholds and accounting for systematic and experimental uncertainties are crucial improvements needed to fully realize the method's potential in practical applications. An important constraint of our current implementation is that while correlations between object vector components are preserved, the model lacks access to event-wise information which impacts the reconstruction accuracy of certain event-level observables. We leave these improvements for future work.

To conclude, our results confirm the versatility of the generalized cDDPM unfolder across diverse physics processes. This non-iterative and flexible posterior sampling approach exhibits a strong inductive bias that allows the cDDPM to generalize to unseen processes without explicitly assuming the underlying distribution, setting it apart from other unfolding techniques so far.

References

[1] Volker Blobel. An unfolding method for high energy physics experiments, 2002. URL https://arxiv.org/abs/hep-ex/0208022.

[2] Anders Andreassen, Patrick T. Komiske, Eric M. Metodiev, Benjamin Nachman, and Jesse Thaler. Omnifold: A method to simultaneously unfold all observables. *Phys. Rev. Lett.*, 124: 182001, 5 2020. doi: 10.1103/PhysRevLett.124.182001. URL https://link.aps.org/doi/10.1103/PhysRevLett.124.182001.

[3] Anders Andreassen, Patrick T. Komiske, Eric M. Metodiev, Benjamin Nachman, Adi Suresh, and Jesse Thaler. Scaffolding simulations with deep learning for high-dimensional deconvolution, 2021. URL https://arxiv.org/abs/2105.04448.

[4] Marco Bellagente, Anja Butter, Gregor Kasieczka, Tilman Plehn, and Ramon Winterhalder. How to gan away detector effects. *SciPost Physics*, 8(4), April 2020. ISSN 2542-4653. doi: 10.21468/scipostphys.8.4.070. URL http://dx.doi.org/10.21468/SciPostPhys.8.4.070.

[5] Kaustuv Datta, Deepak Kar, and Debarati Roy. Unfolding with generative adversarial networks, 2018. URL https://arxiv.org/abs/1806.00433.

[6] Mathias Backes, Anja Butter, Monica Dunford, and Bogdan Malaescu. An unfolding method based on conditional invertible neural networks (cinn) using iterative training, 2024. URL https://arxiv.org/abs/2212.08674.

[7] Marco Bellagente, Anja Butter, Gregor Kasieczka, Tilman Plehn, Armand Rousselot, Ramon Winterhalder, Lynton Ardizzone, and Ullrich Köthe. Invertible networks or partons to detector and back again. *SciPost Physics*, 9(5), November 2020. ISSN 2542-4653. doi: 10.21468/scipostphys.9.5.074. URL http://dx.doi.org/10.21468/SciPostPhys.9.5.074.

[8] Alexander Shmakov, Kevin Greif, Michael Fenton, Aishik Ghosh, Pierre Baldi, and Daniel Whiteson. End-to-end latent variational diffusion models for inverse problems in high energy physics, 2023. URL https://arxiv.org/abs/2305.10399.

[9] Alexander Shmakov, Kevin Greif, Michael James Fenton, Aishik Ghosh, Pierre Baldi, and Daniel Whiteson. Full event particle-level unfolding with variable-length latent variational diffusion, 2024. URL https://arxiv.org/abs/2404.14332.

[10] Sascha Diefenbacher, Guan-Horng Liu, Vinicius Mikuni, Benjamin Nachman, and Weili Nie. Improving generative model-based unfolding with schrödinger bridges, 2023. URL https://arxiv.org/abs/2308.12351.

[11] Anja Butter, Tomas Jezo, Michael Klasen, Mathias Kuschick, Sofia Palacios Schweitzer, and Tilman Plehn. Kicking it off(-shell) with direct diffusion, 2024. URL https://arxiv.org/abs/2311.17175.

[12] Nathan Huetsch, Javier Mariño Villadamigo, Alexander Shmakov, Sascha Diefenbacher, Vinicius Mikuni, Theo Heimel, Michael Fenton, Kevin Greif, Benjamin Nachman, Daniel Whiteson, Anja Butter, and Tilman Plehn. The landscape of unfolding with machine learning, 2024. URL https://arxiv.org/abs/2404.18807.

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

[14] Christian Bierlich, Smita Chakraborty, Nishita Desai, Leif Gellersen, Ilkka Helenius, Philip Ilten, Leif Lönnblad, Stephen Mrenna, Stefan Prestel, Christian T. Preuss, Torbjörn Sjöstrand, Peter Skands, Marius Utheim, and Rob Verheyen. A comprehensive guide to the physics and usage of pythia 8.3, 2022. URL https://arxiv.org/abs/2203.11601.

[15] Cédric Villani. *The Wasserstein distances*, pages 93–111. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. ISBN 978-3-540-71050-9. doi: 10.1007/978-3-540-71050-9_6. URL https://doi.org/10.1007/978-3-540-71050-9_6.

PoS(LHCP2024)234