Machine Learning Methods for Cross Section Measurements

by

Krish Desai

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Physics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Benjamin Nachman, Co-chair
Professor Uros Seljak, Co-chair
Professor Joshua Bloom
Professor Saul Perlmutter

Spring 2025

Machine Learning Methods for Cross Section Measurements

by

Krish Desai

To my family, mentors, and friends

for their unwavering support and encouragement

Contents

List of Figures

List of Tables

Chapter I

Symmetries in data: connections to unfolding challenges.

I.A Symmetries and unfolding.

I.A.1 The complementary nature of symmetry discovery and unfolding.

Unfolding, as discussed, refers to the inverse problem of inferring underlying truth level distributions from observed detector level data, accounting for distortions due to limited resolution and acceptance. Symmetry discovery, aims to identify invariant transformations of the data, that is to say, operations under which the probability distribution of the dataset remains unchanged in a statistical sense [**shaw_lie_2025**, **shaw_symmetry_2024**, **hagemeyer_learning_2022**].

At first glance, these two tasks appear distinct; one concerns recovering numerical distributions, while the other uncovers structural invariances. However, symmetry discovery and unfolding are in fact complementary facets of data driven inference, and integrating the two can yield deeper insights and improved measurements.

From a conceptual standpoint, both tasks share the common goal of revealing hidden truth from observed data. Unfolding endeavours to remove the "detector mask" and recover the true differential cross section or underlying distribution that generated the measurements. Symmetry discovery seeks to reveal underlying structures or

invariances in the data—patterns that persist under transformations, reflecting fundamental symmetries of the physical process or the measurement apparatus.

In practice, these goals can be intertwined. If one discovers a symmetry in the dataset, that knowledge can constrain the unfolding procedure by reducing the effective degrees of freedom in the solution space. Conversely, a properly unfolded distribution is expected to manifest latent symmetries that may have been obscured by detector effects in the raw data. Thus, identifying a symmetry and unfolding a distribution reinforce one another. The former provides a guiding principle or constraint for the latter, while the latter provides a cleaner canvas on which the former can be observed.

Imposing a symmetry as a prior constraint in unfolding can be seen as a form of physically motivated regularisation. For example, architectures that conserve four–momentum or enforce Lorentz invariance by design, as discussed in **??**, effectively impose such constraints, narrowing the set of viable solutions. By restricting solutions to those that respect a discovered or expected invariance, one reduces the space of admissible unfolded distributions to those that are physically plausible, which leads to improved stability and fidelity of the results [**Brehmer:2024yqw**].

This principle has been implicitly utilised in classical unfolding and simulation based calibration. For example, one could assume certain symmetries such as isotropy or detector uniformity when designing the response matrix or when combining

symmetric regions of phase space to reduce uncertainties. With a data driven symmetry discovery tool in hand, one need not rely solely on presumed symmetries. Instead, one can verify them empirically or even discover unexpected invariances. In turn, these empirically verified symmetries can be fed back into the inference pipeline to sharpen measurements, such that a discovered symmetry can inform the unfolding algorithm so that the final measured distribution upholds the invariance.

It is instructive to compare symmetry discovery and unfolding side by side to appreciate their complementary roles. Table I.1 summarises the differences and points of contact between the two.

While unfolding typically requires an explicit model of the measurement process[1] and often relies on supervised learning or iterative inversion techniques, symmetry discovery can be pursued in an unsupervised manner, requiring only the dataset and a class of transformations to probe. The outcome of unfolding is a corrected distribution intended for direct physical interpretation. The outcome of symmetry discovery is a characterisation of invariances, a set of transformations T such that the dataset's distribution is invariant under T within statistical uncertainties. These outcomes are different in nature, but they are mutually beneficial.

Knowledge of invariant structure can guide numerical inference, and conversely, obtaining a more accurate numerical distribution makes it easier to discern subtle

[1] E.g., a response matrix or a parametrised detector simulation.

invariant patterns. In essence, symmetry discovery and unfolding can form a feedback loop in the broader endeavour of measurement and inference, where each can improve the other.

I.A.2 Symmetry aware cross sections.

Since differential cross section measurements lie at the core of particle physics, achieving high precision in these measurements is essential, as even subtle deviations between the measured spectra and theory predictions can signal new physics or the need for refined models. In this context, symmetries play a pivotal role in both the design and interpretation of cross section measurements.

Many physical processes come with known symmetry expectations. For instance, in proton–proton collisions producing particle jets, one expects azimuthal symmetry about the beam axis, i.e. the physics is invariant under rotations in the plane perpendicular to the beam. Consequently, the differential cross section should, after correcting for detector non-uniformities, be independent of the absolute azimuthal angle ϕ of a jet or dijet system [**Chen:2025rjc**, **Spinner:2025prg**, **Pedersen:2023fgr**, **Froidevaux2009ExperimentalCollider**].

Likewise, for processes initiated by identical colliding particles, one often anticipates a symmetry between forward and backward directions. In a symmetric

Table I.1: Comparison of unfolding and symmetry discovery approaches in high-energy physics data analysis. While unfolding aims to correct detector effects to recover true physical distributions, symmetry discovery seeks to identify invariance properties directly from data. These complementary techniques can be combined synergistically: unfolding provides cleaner distributions where symmetries become more apparent, while discovered symmetries can constrain and regularise the unfolding process. This bidirectional relationship enables more robust extraction of physical information from experimental data, particularly in scenarios where detector effects might obscure underlying symmetries or where symmetry constraints can help resolve unfolding ambiguities.

Aspect	**Unfolding**	**Symmetry discovery**
Primary goal	Reconstruct true distributions from detector level observations by inverting detector response	Identify transformation groups under which distributions remain invariant
Inputs	• Measured detector level data • Detector response matrix or • Regularisation scheme	• Measured or unfolded data • Parametrised transformations • Invariance test statistics
Outputs	• Differential cross sections • Probability density functions • Uncertainties/correlations	• Symmetry generators/parameters • Invariance confidence levels • Conserved quantities
Basis	Incorporates domain knowledge through priors/constraints to regularise ill posed inverse problem	Tests generic transformation classes allowing data driven discovery of invariances
Approach	Statistical inference problem to invert $p(x) = \int r(x \mid z)\, p(z)\, \mathrm{d}z$	Hypothesis testing framework to evaluate $p(x) \stackrel{?}{=} p(g(x)) \cdot \lvert g'(x) \rvert$
Benefit	Produces distributions where true physical symmetries manifest more clearly, enabling validation of theoretical predictions and discovery of emergent invariances	Discovered symmetries constrain unfolding solution space, reducing regularisation dependence and improving stability of deconvolution

proton–proton collider, this implies that the rapidity distribution of a centrally produced system [2] should be symmetric about zero rapidity [**Cheung:2017loo**, **CMS:2011xqa**, **Cotogno:2020iio**]. Such symmetry means that the cross section for producing a system at rapidity $+y$ is the same as at $-y$, all else being equal. If a measured differential cross section exhibits a significant asymmetry in these variables after unfolding and acceptance corrections, it would either indicate a previously unaccounted detector bias or hint at a physical effect, both of which are of interest to investigate.

Being symmetry aware in a measurement can mean two things. First, verifying that expected symmetries are indeed present, within uncertainties, in the data, and second, leveraging those symmetries to improve the measurement. On the verification side, symmetry considerations provide valuable consistency checks. Experiments can test whether their unfolded distributions respect fundamental symmetries, and a failure to observe an expected symmetry is a red flag, prompting scrutiny of systematic effects or potential new physics contributions.

On the other hand, when a symmetry is confirmed, one can exploit it to gain statistical and systematic advantages. For example, if a distribution is believed to be symmetric in a certain variable, one can “augment" or combine data from symmetric regions, effectively doubling the effective statistics for that distribution. A common

[2] E.g., dijet pair.

practice is to report differential cross sections as a function of $|y|$ or other symmetry reduced variables, which assumes $y \leftrightarrow -y$ symmetry and thereby reduces statistical fluctuations [**CMS:2013zfg**, **ATLAS:2014rjv**, **ATLAS:2016vlf**]. By incorporating symmetry in this manner, uncertainties can be reduced and the measurement becomes more robust against localised fluctuations.

Symmetry aware analysis also serves to impose physically motivated constraints that guard against overfitting noise in the unfolding process. If one has discovered that a distribution must be invariant under a transformation,[3] imposing this invariance in the unfolding procedure will tie neural network parameters or bin values together that would otherwise float independently. This effectively decreases the number of free parameters describing the unfolded result, acting as a regulariser that prefers solutions consistent with the symmetry. The net effect is an improvement in the precision of the measured cross section and a reduction in spurious oscillatory features that might arise from statistical fluctuations. Moreover, by reducing the dependence on bins with low occupancy (because they are combined with their symmetric counterparts), binned symmetry aware unfolding can also mitigate the impact of detector acceptance edges or inefficiencies in specific regions of phase space.

[3] E.g., rotating the entire event by some angle, or exchanging two identical particles in the final state.

It is important to note, however, that any symmetry based constraint should be applied with careful consideration. One must ensure that the symmetry is either theoretically well founded or empirically validated, lest one impose a false invariance and obscure a genuine asymmetry. This caution further motivates the need for data driven symmetry discovery and validation tools.

One can use methods like SymmetryGAN [**PhysRevD.105.096031**] to verify whether the data uphold the symmetry to a high degree of confidence. Only then would one proceed to incorporate that symmetry into the unfolding process or in the presentation of results. Thus symmetry aware differential cross section measurements can harness known, or discovered invariances to enhance precision and reliability, while simultaneously providing a framework to detect symmetry violations that could point to new phenomena.

The research presented in this chapter builds upon the foundations laid in earlier chapters of this thesis, extending the paradigm of symmetry utilization in the context of measurement and unfolding. **??**, in its survey of existing techniques, provides a statistical foundation for incorporating known symmetries into data analyses. It also highlights the challenges such approaches face, such as the need for careful validation of assumptions. **??** includes references to how *a priori* known symmetries can be hard coded into machine learning models, introducing Lorentz group equivariant networks that guarantee Lorentz invariance in the unfolding of particle physics

data [**Bogatskiy:2020tje**]. The inclusion of symmetry constraints in unfolding models described in **??????** through preserving physical invariants like momentum or charge conservation in the generative model would reduce the solution space and lead to more physically plausible unfolded results.

All of these strategies rely on prior knowledge of the symmetry. This chapter shifts to an data driven approach that provides a novel, flexible, and fully differentiable deep learning based method to discover symmetries directly from data using adversarial learning, which then might allow one to leverage those discovered symmetries to inform the unfolding process. This perspective emphasises the overarching theme of the thesis, the interplay of measurement and inference, by using data driven insights in the form of symmetry discovery to enhance the core measurement task of differential cross section unfolding.

The remainder of this chapter is organised as follows. Section I.B and **??** introduce the formal statistical definition of a dataset symmetry, addressing subtleties like Jacobian volume effects via the concept of an inertial reference density. THey also provides a brief overview of the symmetries most relevant to HEP. Section I.D presents the SymmetryGAN framework, which employs a generative adversarial network to automatically learn symmetry transformations from data. Section I.E validates this approach on illustrative examples and then applies SymmetryGAN to simulated dijet events, demonstrating how it can uncover physically meaningful symmetries

in collider data. Section I.F discusses how the discovered symmetry information can be used to constrain unfolding problems: we outline methods to incorporate symmetry constraints into the unfolding procedure to reduce uncertainties and bias. In Section I.G, the chapter introduces a symmetry aware unfolding methodology and discusses how enforcing the symmetries identified by SymmetryGAN can improve the precision of differential cross section measurements. Finally, the chapter concludes by highlighting how the insights gained here connect back to the broader narrative of the thesis, reinforcing the benefits of combining machine learning driven discovery with principled measurement techniques.

I.B Formalism and Importance

In physics and statistics, a symmetry refers to an invariance of a system or dataset under a well-defined transformation. Formally, let G be a group of transformations (continuous or discrete) acting on a space of states or observations X. A physical system or probability distribution is symmetric under G if applying any transformation $g \in G$ leaves the relevant observables unchanged. In group theoretic terms, there exists an action $g : x \mapsto g(x)$ such that for all $x \in X$ and $g \in G$, the value of an function $f(x)$ remains equal to $f(g(x))$. However, if $p(x)$ denotes a probability density on X, a group G is a symmetry of p if

$$\forall g \in G \; p(x) \; \mathrm{d}x = p(g(x)) \; \mathrm{d}(g(x)) \, . \tag{I.1}$$

In measure theoretic language, a symmetry corresponds to an invariant measure. For any measurable subset $A \subseteq X$ and any transformation $g \in F$, g is a symmetry of A if $\mu(A) = \mu(g \cdot A)$, meaning the measure assigned to outcomes in A is the same as that for the transformed set $g \cdot A$. This definition encompasses both continuous symmetries[4] and discrete symmetries [5]. Symmetry principles lie at the heart of modern particle physics and also strongly influence experimental measurements. At the theoretical level, fundamental symmetries constrain the form of physical laws

[4]Lie groups, such as rotations depending on a continuous angle parameter.

[5]groups constructed as Jordan–Hölder extensions [**solomon_brief_2001**, **HolderDieOrdnungszahlen**, **jordan_traite_nodate**] e.g. a mirror reflection or a permutation of identical objects

and often correspond to conserved quantities or selection rules. At the data level, symmetries, and their breaking, shape the distributions of observed events and can be leveraged for more efficient data analysis. In the context of colliders, many observables are governed by symmetries of the underlying theory as well as symmetries introduced by the detector and measurement process. It is therefore crucial to articulate how these symmetries operate both in ideal physics scenarios and in real observations. This section provides a rigorous overview of symmetries relevant to collider physics and measurements. It begins with the fundamental symmetries in particle physics that underlie observable phenomena in Section I.B.1. Section I.B.2 discusses symmetries in detector response functions and how the measurement apparatus can preserve or violate underlying invariances. Next, Section I.B.3 examines how symmetries manifest in measured cross sections and data distributions, clarifying the translation from physical symmetry to statistical patterns in experimental histograms. Finally, Section I.B.4 highlights the challenges in identifying symmetries from noisy data, setting the stage for data–driven symmetry discovery techniques. This foundation will be essential for later sections that introduce methods like SymmetryGAN for learning symmetries from data.

I.B.1 Fundamental symmetries in HEP.

Particle physics is built upon a framework of symmetries that determine the allowed forms of interactions and the conservation laws observed in experiments. Spacetime symmetries, in particular subgroups of the Poincaré group, are foundational. The Poincaré group includes continuous Lorentz invariance (rotations and boosts) and translations (in spacetime). Lorentz invariance implies that the laws of physics take the same form in any inertial reference frame. Equivalently, physical observables can be expressed in terms of Lorentz invariant quantities[6] that remain unchanged under boosts or rotations. For example, the Mandelstam variables s, t, u in a scattering process or the decay angle distribution in a particle's rest frame are formulated to respect Lorentz symmetry. In practice, exact Lorentz invariance means there is no preferred direction or absolute velocity in the underlying theory. A given interaction process should yield identical outcomes whether the laboratory frame is, say, Earth bound or boosted to a constant velocity. As a consequence of Lorentz symmetry and spatial isotropy, angular momentum and linear momentum are conserved in isolated systems[7]. Time translation symmetry leads to energy conservation, ensuring that system's total energy and the collision centre of mass energy are fixed constants of motion. These spacetime symmetries are exact symmetries of all known fundamental

[6] E.g., invariant masses, angles, and dimensionless ratios.

[7] Noether's theorem associates these conservations with rotational and translational symmetry, respectively [**noauthor_nachrichten_nodate**]

interactions and provide the basis for defining covariant formalisms in quantum field theory.

Beyond spacetime, the internal symmetries of the Standard Model dictate the spectrum of particles and their interactions. Chief among these is the gauge symmetry group $SU(3)_C \times SU(2)_L \times U(1)_Y$, which defines quantum chromodynamics and electroweak theory. Gauge symmetries are local symmetries that require the introduction of gauge bosons; although these are internal symmetries rather than symmetries of observable spacetime, they have observable consequences such as electric charge conservation, associated with $U(1)_Y$ hypercharge symmetry, and the existence of multiple particle generations. The gauge symmetries of the Standard Model are spontaneously broken in certain cases.[8] However, even broken symmetries leave remnant effects, such as the custodial symmetry in the Higgs sector or approximate conservation of isospin in QCD. These internal symmetries set selection rules. Processes that violate gauge charge conservation are forbidden and decays proceed only via symmetric channels.

Alongside continuous symmetries, several discrete symmetries play a crucial role in particle physics. The most prominent are C (charge conjugation, exchanging particles with their antiparticles), P (parity, spatial inversion or mirror reflection), and T (time reversal). Each of these can be considered a transformation that might leave the

[8] One of the most notable examples is the electroweak $SU(2)_L \times U(1)Y$ breaking to $U(1)_{\rm EM}$ via the Higgs mechanism, which introduces masses for the $W^{\pm}$ and Z bosons and differentiates the electromagnetic and weak interactions.

fundamental laws invariant. In the Standard Model, CP symmetry is approximately a symmetry of electromagnetic and strong interactions, but notably broken in weak interactions. This manifests as differences in the behaviour of matter and antimatter. Like most notable instance of this is the well known CP violation in neutral kaon and B-meson decays means those processes occur at different rates or with different phase relationships than their CP-mirrored counterparts [**Neubert:1996qg**]. If CP were an exact symmetry of the dynamics, one would expect, for example, the angular distribution of decay products in a mirror–reflected process, swapping particles for antiparticles, to be identical to the original. The observed deviations are vital clues to physics beyond simple symmetries. Parity by itself is also violated maximally in the weak interaction.[9] On the other hand, the strong and electromagnetic interactions conserve parity, so for many processes, especially at high energies where electroweak effects are subdominant, it is a good symmetry. A process governed by QCD, like multijet production, should occur equally in a configuration and its mirror reflected image, unless the experimental setup selects a handedness. Charge conjugation is likewise not a symmetry of the full Standard Model, since, for example, there are no right handed neutrinos to pair with left handed ones under C, but for purely electromagnetic processes C symmetry implies, that producing a negatively charged

[9] classic examples are the left handed nature of neutrinos and the parity asymmetric angular distribution of electrons in polarized ^{60}Co beta decay [**Wu:1957my**]

particle is as likely as producing the corresponding positively charged antiparticle under equivalent conditions.

Importantly, the combination CPT is believed to be an exact symmetry of local quantum field theory. CPT symmetry implies, for instance, that particle and antiparticle masses and lifetimes are exactly equal [**Kostelecky:1998ic**]. While CPT is typically not directly tested by single distribution symmetries in HEP experiments, it provides a fundamental consistency check on any observed CP or T violation. Table I.2 summarises these fundamental symmetries, their group theoretic character, and their status in the Standard Model.

Table I.2: Summary of fundamental symmetries relevant to particle physics, their group theoretic structure, and experimental signatures. The table includes spacetime symmetries, gauge symmetries, discrete symmetries, and quantum statistical symmetries. "Conserved charge" refers broadly to conserved quantities arising from continuous symmetries via Noether's theorem or to quantum numbers constrained by discrete symmetries. The "Status in SM" column indicates whether each symmetry is exact, approximate, or explicitly broken within the Standard Model. Observable signatures provide experimental handles for testing these symmetries. The hierarchy of symmetry breaking guides the search for BSM physics through precision tests of invariance.

Symmetry	**Group structure**	**Conserved charge**	**Observable signatures**	**Status in SM**
Lorentz-invariance	$\mathrm{SO}(3,1)$	$x^\mu p^\nu - x^\nu p^\mu$	Invariant mass; angular distributions	Exact [**Kostelecky:2010ze**]
Spacetime translation	$\mathbb{R}^{1,3}$	p^μ	Missing momentum; vertex momentum conservation	Exact [**Weinberg:1995mt**]

Continued on next page

Table I.2 continued from previous page

Symmetry	Group structure	Conserved charge	Observable signatures	Status in SM
Gauge-symmetry	$SU(3)_C \times SU(2)_L \times U(1)_Y$	Colour, isospin, hypercharge	Jet colour flow patterns; W boson charge asymmetry;	Exact (local) [**peskin_introduction_1995**]
Electroweak	$\langle H \rangle \neq 0$	M_W, M_Z	$\rho = M_W^2/(M_Z^2 \cos^2\theta_W)$	Broken [**PhysRevLett.13.508**]
Charge conjugation (C)	$\mathbb{Z}_2 : \psi \to \mathcal{C}\bar{\psi}^T$	$\alpha_{q\pm} \leftrightarrow \alpha_{\bar{q}\mp}$	e^+/e^- production ratios; $\pi^0 \to \gamma\gamma$, $\pi^0 \not\to 3\gamma$	Conserved in QED/QCD; violated in weak [**Wu:1957my**]

Continued on next page

Table I.2 continued from previous page

Symmetry	Group structure	Conserved charge	Observable signatures	Status in SM
Parity (P)	$\mathbb{Z}_2$: $\boldsymbol{x} \to -\boldsymbol{x}$	$\eta_P = \pm 1$	Neutrino helicity; asymmetric β decay	Conserved in QED/QCD; violated in weak [**Lee:1956qn**]
CP	$\mathbb{Z}_2 \times \mathbb{Z}_2$	$\delta_{CP}, \theta_{\text{strong CP}}$	B^0-$\bar{B}^0$ mixing asymmetry; kaon decay: $\epsilon_K \sim 10^{-3}$	Violated: $\delta_{CP} \approx 68\,\text{deg}$ [**Charles2005CKMFactories**]
Time reversal (T)	$\mathbb{Z}_2 : t \to -t$	Amplitudes	Electric dipole moments; $K_L \to \pi^+\pi^- e^+ e^-$ decay	Violated (CPT theorem) [**Luders:1957zz**]

Continued on next page

Table I.2 continued from previous page

Symmetry	**Group structure**	**Conserved charge**	**Observable signatures**	**Status in SM**
CPT	$\mathbb{Z}_2$ (anti-unitary)	$m, \tau_{q\pm} = m, \tau_{\bar{q}\mp}$	$\lvert m_p - m_{\bar{p}}\rvert/m_p < 10^{-10}$; $\lvert\tau_\mu - \tau_{\bar{\mu}}\rvert/\tau_\mu < 10^{-5}$	Exact (theorem) **[Streater:1989vi]**
Permutation	S_n/A_n	Bose/Fermi stats	HBT correlations in $\pi^+\pi^+$; Pauli exclusion in spectra	Exact **[Pauli:1940zz]**

Beyond the Standard Model's built in symmetries, there are approximate global symmetries that often prove useful in particle physics. Examples include isospin symmetry, an $SU(2)$ symmetry treating up and down quarks as identical in the limit of equal masses, and flavour symmetries, like the $SU(3)$ of the light *uds* quarks, both of which are not exact, but underlie patterns in hadron production and decay. For instance, isospin symmetry implies that processes differing only by swapping an up quark with a down quark, such as producing a proton versus a neutron, have nearly equal cross sections, up to corrections from the up–down mass difference or electromagnetic effects.

Similarly, the universality of physical laws under interchange of identical particles leads to permutation symmetry. If two particles of the same type appear in a final state, the probability distribution is invariant under exchanging them. In quantum theories this is enforced by (anti)symmetrization of identical particle states. In HEP observables, permutation symmetry means that one cannot physically distinguish, say, which of two identical jets in an event is 'jet 1' or 'jet 2'—any labelling is arbitrary and the underlying physics treats the two jets on equal footing. When calculating cross sections, this symmetry is accounted for by dividing by the a symmetry factor to avoid over counting identical configurations.

In practical analyses analysis one often has to impose an ordering, such as 'leading' and 'subleading' jet by momentum, for convenience, but the fundamental permutation

invariance implies that any physical conclusion should not depend on this arbitrary ordering.

In summary, fundamental symmetries, Poincaré (Lorentz and translations), gauge invariances, discrete symmetries like CPT, and permutation invariance for identical particles, provide a set of invariance principles for particle interactions. These symmetries constrain the form of theoretical cross sections and transition rates.

Many measurable quantities in experiments such as cross sections, angular distributions, etc., either reflect these symmetries, when they hold, or provide avenues to detect symmetry breaking when deviations are observed. However, the symmetries of nature at the fundamental level are not always manifest in what detectors actually record. We next turn to how the detector response and measurement process can modify or obscure these symmetries.

I.B.2 Symmetries in detector response functions

The detector response function $r(x \mid z)$ describes the probability of observing a measurement outcome x given a true state z. An ideal detector with perfect resolution would preserve all physical symmetries present at the particle level. In reality, detectors often break or reduce symmetries that the underlying physics possesses. Understanding which symmetries are preserved, approximated, or lost

convolution with $r(x \mid z)$ is crucial for interpreting measured data. Here we discuss several common invariances and how they are affected by realistic detectors in their response.

I.B.2.i Spatial uniformity and rotational symmetry.

Many detectors are designed with a roughly cylindrical geometry around a natural axis, such as the beam axis in collider experiments, aiming for azimuthal symmetry. Hence they are designed to provide close to uniform coverage in the plane normal to the axis.

Ideally, if the physical process yields a uniform distribution in the azimuthal angle ϕ (i.e. no preferred direction around the axial line), a perfectly symmetric detector would register an equal number of events in each azimuthal segment. In practice, certain asymmetries are present in any measurement.

For example, a detector may have support structures or cabling at certain angles, or irregular segmentation, leading to variation in efficiency with ϕ. The electromagnetic calorimeter (ECAL), as an illustration, might be segmented into modules that cover specific ϕ slices. Given this, events falling into the gap between modules could be recorded with lower efficiency or energy resolution, creating a $\phi-$dependence in the observed data even if the true distribution was uniform. Detectors often have periodic segmentation, meaning continuous rotational invariance is broken down to

a discrete rotation symmetry, invariant only under rotations corresponding to full module spacings.

As a concrete example, imagine a detector with 360 identical modules each covering $\Delta\phi = 1°$. This detector is invariant under rotation by multiples of 1 deg increments, but a rotation by an arbitrary angle (say 0.5 deg) would lead to a different alignment of a particle's trajectory with respect to module boundaries, yielding a measurably different response [**Nabat:2024nce**]. Thus, the continuous symmetry is reduced to a discrete one, and even that discrete symmetry may be imperfect if modules are not exactly identical or have time varying efficiency.

Thus azimuthal symmetry at the physics level is usually preserved approximately by detector design, but slight non-uniformities in ϕ response are common and must be accounted for either via calibration or acceptance corrections.

I.B.2.ii Polar coverage and boost invariance.

Often detectors in HEP experiments, especially collider experiments, also have limited coverage in the polar direction, along the beam axis. No real detector covers the full 4π solid angle; there is always a cut off at some polar angle (or pseudorange η) beyond which particles escape detection. In particular, the forward regions close to a beam are notoriously difficult to instrument.

This breaks the full spherical symmetry of space. A process that is symmetric under arbitrary rotations, such as a perfectly isotropic decay in its rest frame, will not appear isotropic in the laboratory measurement if a significant portion of the solid angle is unobserved. Detectors are typically ideally symmetric under rotations about the beam axis but not under arbitrary rotations that tilt the beam axis, because the beam direction is a fixed axis of symmetry.

In other words, the presence of beams singles out a preferred direction, the beam axis $\hat{z}$, and detectors are built around this axis. Consequently, the data may reflect cylindrical symmetry, invariant under $SO(2)$ rotations around $\hat{z}$, but not full $SO(3)$ rotational symmetry.

This also has consequences for Lorentz boost invariance. While the underlying physics is Lorentz invariant, the detector is a fixed apparatus in one frame. A boost along the beam direction, i.e. a change to the reference frame moving with respect to the collision, will generally change how events are distributed relative to the detector acceptance.

As an example, consider a boost that causes particles to have higher longitudinal momentum; in the lab frame, more particles will end up at small polar angles, closer to the beam line, where detection efficiency is lower, thus the observed distribution of, say, pseudorapidity η will shift. The detector has a finite acceptance in η, so a Lorentz boost that moves events into the far forward region will result in a fraction

of events being lost. Therefore, the measured distributions are not invariant under Lorentz boosts, even though the underlying parton level kinematics can be expressed in Lorentz invariant terms.

Thus the physical construction of detectors break global translational and boost symmetry by virtue of being static and having edges. A high energy interaction viewed in different inertial frames is physically identical, but a detector at rest in the lab frame will record it differently unless one corrects for acceptance and inefficiencies.

I.B.2.iii Resolution effects and approximate invariance.

Even if a symmetry could hold in principle, the resolution and threshold effects of detectors often spoil exact invariance. A salient example arises with Lorentz invariance and invariant mass reconstruction.

As a thought experiment, imagine a two body decay producing a pair of muons, such as $Z^0 \to \mu^+\mu^-$. The true invariant mass of the muon pair is fixed, irrespective of the Z boson's momentum, because it is simply a Lorentz scalar. However, a detector measures muon momenta with finite precision, and that precision typically degrades at high momentum[10]. If the Z boson is produced nearly at rest in the lab, its decay muons have moderate momenta and the detector might reconstruct the invariant mass with a narrow resolution. If instead the Z is produced with a large boost, the

[10]Tracking detectors determine momentum from curvature in a magnetic field, which becomes very small for high momentum muons, leading to larger relative uncertainty in the measurement.

muons each have higher lab frame momenta, and the detector's momentum resolution broadens the reconstructed mass distribution. The result is that the distribution of reconstructed $m_{\mu\mu}$ for boosted Z events is broader (and potentially biased) compared to that for non-boosted events.

Thus, a Lorentz boost, which should not matter to an ideal measurement, actually changes the statistical distribution of an observable due to detector response. This is an example of an approximately respected symmetry. At low boost the symmetry holds well, but at high boost the symmetry is effectively broken by detector effects.

Similarly, thresholds in detector sensitivity (e.g. a calorimeter that only records energy above some minimum) can break symmetry under transformations that redistribute energy. A detector that is equally efficient for electrons and positrons should demonstrate C symmetry. However, if the process produces a broad energy spectrum, imposing a cut on low energy particles can introduce a bias. Low energy e^+ are more likely to be lost than e^- due to their different interaction rates with detector material. In such cases, even an underlying physical symmetry may not yield equal measured counts.

I.B.2.iv Mirror and charge symmetry.

Detectors are not usually built to be fully symmetric under parity inversion or charge conjugation, even though one often assume these symmetries for the relevant

physics should reflect in data. A parity inversion would swap the 'forward' and 'backward' directions in the detector. If the detector has identical coverage in the forward ($+z$) and backward ($-z$) hemispheres, one could say that it is parity symmetric with respect to the interaction point.

Many detectors strive for this by having symmetric endcaps on both sides of the interaction region. However, even then, subtle asymmetries can exist because it is not feasible to prevent one side from having a slightly different material distribution or a different calibration from the other. As a result, a process that is forward–backward symmetric in its physics[11] might show a forward–backward asymmetry in the raw data. Experiments typically correct for such differences by equalizing calibrations, but the point stands that the intrinsic detector response can break the symmetry.

Likewise, charge conjugation symmetry in detection would mean the detector is equally sensitive to positive and negative charges. While the detector electronics and geometry generally don't prefer one charge sign, magnetic fields introduce a notable asymmetry, because charged particles bend in opposite directions, and this can lead to charge dependent acceptance. In a magnetic spectrometer, positively charged particles bend in one direction and negatively charged particle in the opposite direction. If the acceptance boundaries, like the edge of the detector volume, cut off

[11] In $p-p$ collisions for example, the two beam directions are equivalent so the distribution of particles as a function of rapidity y should be symmetric about $y = 0$.

tracks in one curvature direction more than the other, one will observe a difference in detection rates even if production is symmetric.

Another example is that the different interaction of e^+ and e^- with matter could lead to different detection efficiencies. These are lower order effects, but they illustrate that a detector is a physical object that need not respect the abstract symmetries of the theory. Careful simulations and calibrations are performed to quantify and mitigate these asymmetries in HEP experiments.

I.B.2.v Permutation symmetry and identical particles.

Physically, as noted, swapping two identical particles should change nothing in an ideal measurement. Detectors, however, could introduce differences. Two identical particles (say two photons) that go into different regions of the detector can have their energies might be measured with different resolutions or one might pass quality cuts and the other fail due to region specific noise. As a result, the joint distribution of the two particle system in the measured data might not be symmetric under exchange, even though it was at truth level.

As a another simple example, consider two jets in an event for which, at the particle level, the probability $P(E_1, \eta_1; E_2, \eta_2)$ is symmetric under $(1 \leftrightarrow 2)$. After detection, suppose that jet 1 falls in the central barrel, with excellent energy resolution and jet 2 falls in the forward region, with poorer resolution and lower efficiency. The

measured energies $E_1^{(\text{meas})}$ and $E_2^{(\text{meas})}$ will have different response smearing. If one then orders jets by measured energy and calls the highest energy jet the leading jet, the distribution of leading and subleading jets will not mirror one another exactly. Effectively, the detector induced asymmetry has assigned labels to the jets where none existed. Analysts must be wary of these effects; one option used is to 'symmetrize' the analysis if possible to recover the permutation symmetry that the physics assures.

Table I.3 summarises a few key examples of how an ideal symmetry at the particle level can be broken or reduced by detector effects. These examples illustrate why fully accounting for detector response is essential when testing physical symmetry hypotheses with data.

Despite these challenges, experimentalists strive to design detectors with as much symmetry as feasible and to correct for known asymmetries. For instance, collider detectors often have nearly full 2π azimuthal coverage and layered symmetries, segmenting in ϕ and η uniformly, specifically to preserve rotational invariance and facilitate combining data over symmetric regions. Detector simulation and calibration are used to quantify symmetry breaking. If a $\phi-$dependence is observed in calibration data, it can be corrected so that the final analysis treats those variations as a systematic uncertainty or removes them. Nonetheless, the reality remains that physical symmetries can fail to translate into measured symmetries.

Table I.3: Ideal SM symmetries and detector induced symmetry breaking in HEP experiments. While fundamental interactions may respect certain symmetries exactly, detector geometries, material distributions and reconstruction algorithms violate these symmetries at measurement level. The table illustrates how common detector limitations transform SM symmetries. These effects must be carefully modelled in simulations and corrected through calibration to extract the underlying physics.

Symmetry	**Ideal outcome**	**Detector effect**
Azimuthal rotation $\phi \to \phi + \alpha$	Rotational invariance: uniform distribution in $\phi \in [0, 2\pi]$. No preferred transverse direction	Discrete n-fold symmetry: detector segmentation creates $\phi-$dependent acceptance. Dead material between modules introduces periodic inefficiencies. Triggers based on detector regions further break symmetry

Continued on next page

Table I.3 continued from previous page

Symmetry	**Ideal outcome**	**Detector effect**
Lorentz boost $\beta_z \to \beta_z + \Delta\beta$	Physics invariant under longitudinal boosts; cross sections expressible in terms of Lorentz scalars.	Lab frame dependence: fixed detector geometry defines preferred frame. Forward boosts push particles beyond acceptance. Resolution degrades for highly boosted objects. Trigger thresholds defined in lab frame are not invariant
Parity $\mathbf{x} \to -\mathbf{x}$	For P conserving processes: equal rates and distributions for original and parity transformed events	Geometric asymmetry: detectors rarely possess reflection symmetry about $z = 0$. Forward and backward regions have different instrumentation, acceptance, and resolution. Magnetic field direction picks out handedness

Continued on next page

Table I.3 continued from previous page

Symmetry	**Ideal outcome**	**Detector effect**
Charge conjugation $q \to -\bar{q}$	C symmetric interactions produce particles and antiparticles with identical rates and kinematic distributions	Charge dependent efficiency: opposite charges bend oppositely in magnetic field, sampling different detector regions. Material interactions differ (e.g., K^+ vs K^- nuclear cross sections). Trigger and particle ID algorithms may have charge bias
Permutation Symmetry $(i \leftrightarrow j)$	For identical particles: joint distribution exhibits exchange symmetry	Position dependent response: particles in different regions experience different resolutions, efficiencies, and systematics. Ordering by p_T obscures underlying symmetry. Combinatorial background differs for same region vs different region pairs

In the language of probability distributions, if $p(z)$ is invariant under transformation T, but the response $r(x \mid z)$ is not invariant in the corresponding way, then the folded distribution $p(x) = \int r(x \mid z)\, p(z), \mathrm{d}z$ will not be invariant under T applied to x. Only if both p_{truth} and r share the symmetry T will p_{data} exhibit it. This

conceptual understanding is vital when one interprets measured cross sections and tries to infer or discover symmetries from data.

I.B.3 How symmetries manifest in measured cross sections.

Given the above considerations, one can now examine how symmetries and symmetry violations are reflected in the measured distributions that experiments record and report. A measured cross section differential in some observable is effectively a statistical aggregate of many collision events, after selection cuts and corrections. If the underlying physics possesses a symmetry, one might expect the differential cross section to reflect that, provided the measurement process does not hide or distort it. In practice, one observes in histograms a mixture of genuine physical symmetry patterns and effects of detector acceptance or sample selection.

I.B.3.i Exact symmetries and flat distributions.

A hallmark of a symmetry in a distribution is a repeated pattern indicating invariance. For example, consider azimuthal invariance in a proton–proton collision. Since the colliding protons provide a cylindrically symmetric initial state of two identical beams head on, no physics process at the parton level prefers a particular ϕ direction. Consequently, the true differential cross section $\mathrm{d}\sigma/\mathrm{d}\phi$ for an inclusive process is invariant under the transformation $\phi \mapsto \phi + \delta\phi$ (aside from small QED

effects or residual detector magnetization influences). If the detector has uniform ϕ coverage and the analysis has no $\phi-$dependent cuts, the measured distribution of events as a function of ϕ should be approximately flat. Any significant deviation from flatness might indicate an instrumental problem or a selection bias.

Once the symmetry has been established, one might combine data from all ϕ slices (since they are equivalent) to improve statistical precision, effectively using the symmetry to gather more data. However, as noted, small modulations can appear if certain detector modules deviate from the rest; these are corrected or quoted as systematic uncertainties.

Another example is rapidities in symmetric collisions: in a $p-p$ collider, for example, at equal beam energies, the centre of mass frame coincides with the lab frame, and the process is symmetric under exchanging the two beam directions. This implies that the distribution of particles in rapidity y is symmetric about $y = 0$ for processes that do not involve a bias.[12] This is why measurements of inclusive jet or hadron yields often present results as a function of $|y|$ or $|\eta|$, the absolute value of rapidity or pseudorapidity, invoking the symmetry $y \leftrightarrow -y$ to double the statistics and simplify presentation. The physical symmetry (identical proton beams) justifies this, and one checks that, within uncertainties, the $+y$ and $-y$ distributions are

[12]For instance, pure QCD dijet production should yield a symmetric $\mathrm{d}\sigma/\mathrm{d}y$ for jets, with equal activity in the forward ($+y$) and backward ($-y$) hemispheres.

consistent before merging. Thus, a symmetry in initial conditions and dynamics (here, invariance under $y \to -y$) leads to a clear symmetry in the measured cross section (equal yields for $\pm y$).

I.B.3.ii Symmetries in kinematic shapes.

Symmetries often impose recognizable shapes or constraints on distributions. For example, energy and momentum conservation require that for each event, the vector sum of momenta of final state particles equals that of initial state. As a result, distributions of total transverse momentum in events would be expected to peak at zero, and any significant imbalance indicates e.g. neutrinos or detector holes. This is not a symmetry in the sense of a group acting on one event's space, but rather a deterministic constraint on the ensemble. The distribution of missing momentum should be centered at zero and isotropic in azimuth. Experiments can thus verify that the missing transverse momentum vector has no preferred direction to validate rotational symmetry and momentum conservation in aggregate.

If one measures the transverse momentum spectrum of the first jet against that of the second jet in dijet events (with jets ordered by p_T), there is no fundamental reason for these spectra to differ except for the ordering bias. The leading jet p_T distribution will be harder by construction, and the subleading softer, but any jet is equally likely to be at a given p_T as its partner, aside from that ordering. This

symmetry can therefore by verified through the similarity between the distribution of subleading jet p_T and the leading jet p_T distribution of a lower energy subset, or by symmetrising the dataset by swapping jets event by event and seeing no change in overall two jet correlation distributions.

In summary, wherever a symmetry exists, one finds redundancies or equalities in the measured spectra: sections of phase space that should mirror other sections. Experimental analyses often exploit this to measure detector backgrounds, by assuming that an uninstrumented region should have similar counts as a well instrumented region after normalisation.

I.B.3.iii Interplay of physical and detector symmetries.

It is important to disentangle which symmetries in a measured cross section come from physics and which from measurement procedure. An analysis might impose a cut that itself introduces a symmetry or asymmetry. When presenting a measured cross section, unfolding detector effects to reconstruct particle level distributions to the extent possible, to report a cross section as it would appear with an ideal detector, can restore the symmetries that belong to the physics by removing the distortions of measurement [**DAgostini:265717**]. For example, if the raw data show a $\phi-$dependence due to detector inefficiency, the unfolded cross section as a function of ϕ should be flat, with larger uncertainties reflecting the correction.

In this sense, symmetries provide a consistency check. If after unfolding one still sees a symmetry violation in a quantity that should be symmetric by physics, the unfolding procedure might be flawed. Conversely, if a symmetry is expected to be broken by physics, one must be careful to ensure the detector is not distorting the size of the asymmetry. For instance, measuring a forward–backward asymmetry in top quark production requires excellent control of any detector differences between the forward and backward directions so that the observed asymmetry can be trusted as a physical sign of potential weak interaction interference (or potentially new physics).

A physical symmetry is a property of the underlying probability law. It requires equal probabilities for events and their transformed versions. A statistical symmetry of a dataset entails that the finite sample of observed data appears invariant under some transformation, within the limits of noise, so that with large data, one expects the symmetry to become apparent as the fluctuations average out. If deviations persist significantly beyond expected fluctuations, that flags either a real symmetry violation or unaccounted systematics.

I.B.4 Challenges in identifying symmetries from noisy data.

Identifying symmetries in experimental data is not always straightforward. Noisy data, stemming from finite statistics, background processes, and detector imper-

fections, can obscure or mimic symmetry signals. This section outlines the main challenges one faces in discerning true invariances or symmetry violations within collider datasets, and the need for methods like the SymmetryGAN approach developed later in this work to address these challenges.

A fundamental challenge is that any empirical distribution has random fluctuations. If an underlying distribution is perfectly symmetric (say truly uniform in ϕ), a finite sample will still exhibit some variation across ϕ bins. Hence any symmetry discovory method must have a mechanism to distinguish a real asymmetry from a mere fluctuation.

Conversely, an underlying asymmetry can be washed out by limited statistics. This is especially pertinent in searches for new symmetries or violations. The signals are often at the level of small deviations and can be difficult to detect over statistical fluctuation. Moreover, multiple comparisons increase the probability that one finds an apparent "symmetric pattern" in some projection of the data purely by chance.

As discussed, detector effects can induce or conceal asymmetries. Often the largest uncertainties in measuring symmetry come from how well we understand the detector. For example, in measuring a forward–backward asymmetry, uncertainties in the relative efficiency of the forward and backward region directly translate to uncertainty in the asymmetry observable. If those efficiencies are poorly known, one might not be able to distinguish a symmetry violation from detector effects.

Similarly, backgrounds and other processes that mimic the signal might not share the symmetry of the signal. Suppose one is looking for a symmetry in a certain particle decay distribution; if there is a significant background from a different process that does not respect that symmetry, the combined data will appear to break the symmetry even if the signal alone is symmetric. Careful background subtraction or isolation is required. In practice, identifying a symmetry often involves comparing two distributions (e.g. $P(x)$ vs $P(Tx)$ for some transformation T) and seeing if they differ. If they do, one must estimate if the difference is due to known systematic effects. This typically demands high-precision calibration. For instance, to confirm CP symmetry in production of particle vs antiparticle to the 10^{-3} level, one needs detector efficiencies known to better than 0.1% between positively and negatively charged particle detection [**Gordon:2013eha**]

I.B.4.i Dimensionality challenges.

HEP interaction events are often high dimensional, consisting of many particles with various kinematic attributes. A symmetry might not be evident in any single one dimensional projection, but rather in a complicated combination of variables. For example, Lorentz invariance is best seen when considering all four–momenta together or invariants like masses; a naive look at just one momentum component would not show it. Permutation symmetry in a multijet event is a property of the

joint distribution of all jet momenta, not necessarily obvious if one only looks at single–jet spectra.

This is where machine learning methods become attractive, because they can, in principle, detect subtle patterns in high dimensional data. However, even ML models need guidance. The space of possible transformations is huge, and hence searching it naively for invariances is intractable. Hence traditional methods often restrict attention to physically motivated symmetry transformations (rotations, reflections, boosts, particle exchanges, etc.).

Since scanning for symmetries by comparing all possible pairs of transformed distributions is computationally prohibitive, as data volumes grow and analysis spaces become more complex, we need more automated symmetry discovery mechanisms. The SymmetryGAN approach discussed in this thesis is one attempt to automate the discovery of symmetries by leveraging generative adversarial networks. Conceptually, SymmetryGAN will trains a generator (applying candidate transformations) against a discriminator to test if the transformed data looks statistically identical to the original data. If the generator generates a transformation under which the discriminator is maximally confounded, that transformation corresponds to a symmetry of the data distribution. Implementing this is challenging: the model must search a continuous space of transformations, handle approximate symmetries, and avoid trivial solutions.

A careful choice of network architecture using known equivariances are needed to make such learning robust.**[cite –KD]**

In summary, identifying symmetries from noisy collider data requires

1. Sufficient statistics and rigorous statistical tests to differentiate real invariances from fluctuations,

2. Precise control of detector systematics to avoid mistaking detector effects for or against symmetry,

3. Methods to probe high-dimensional and subtle symmetry patterns that might elude simple binned analyses, and

4. Methodological consideration to handling approximate symmetries in a principled way.

These challenges motivate the development of tools like SymmetryGAN, which I will introduce in the next sections. Such tools aim to combine physical insight with machine learning's ability to detect patterns, thereby providing a statistical discovery framework for symmetries. SymmetryGAN and similar approaches offer a promising path to unveil symmetries that are latent in complex data. The rigorous understanding of symmetry and symmetry–breaking provided in this section will form the foundation on which those computational methods build, ensuring that

any discovered “symmetry" is physically meaningful and relevant to the challenges of unfolding and analyzing collider data.

I.C Statistical Definition of Dataset Symmetries

The concept of symmetry in physics typically evokes images of rotational invariance in crystals, parity conservation in weak interactions, or gauge transformations in field theory. Yet when we turn our attention to experimental data, especially the high dimensional datasets emerging from modern collider experiments, the notion of symmetry becomes surprisingly subtle. What does it mean for a collection of measured events to possess a symmetry? This question, deceptively simple in appearance, reveals profound connections between statistical inference, group theory, and the fundamental challenge of unfolding detector effects from observed data.

I.C.1 Distinction between point and dataset symmetries

Statistical symmetries in datasets represent a fundamental departure from traditional geometric symmetries, requiring careful consideration of probability measures, transformation Jacobians, and reference densities. This section explores the mathematical foundations, practical applications, and machine learning approaches to understanding and leveraging dataset symmetries. The distinction between symmetries of individual data elements and entire datasets lies at the heart of statistical theory of symmetries. For individual data points, symmetry is characterized by simple invariance: a transformation g preserves element x if $g(x) = x$. This represents a

straightforward geometric notion where specific points remain fixed under transformation. Distribution level symmetries, however, operate on probability measures rather than individual points. A measure space X with probability measure μ exhibits symmetry under group G when the measure remains invariant.[–KD]

$$\forall A \subseteq X \, \forall g \in G \, \mu(A) = \mu(g(A)) \tag{I.2}$$

This measure theoretic definition captures the statistical properties of entire distributions rather than individual elements.

The critical insight, as formalized in the SymmetryGAN framework, is that dataset symmetries are ambiguous due to Jacobian factors introduced during coordinate transformations.[–KD] Unlike point symmetries, where transformations either preserve or don't preserve specific locations, dataset symmetries must account for how probability densities transform under coordinate changes. This fundamental difference necessitates the introduction of inertial reference densities to properly define statistical symmetries.

I.C.2 Inertial reference densities and their theoretical role

The concept of inertial reference densities emerges as a necessary theoretical construct, analogous to inertial frames in classical mechanics.[–KD] These reference densities provide a baseline against which statistical symmetries can be meaningfully

defined, resolving the ambiguity inherent in coordinate transformations of probability measures.

In the formal framework, a reference density $\rho(x)$ establishes a canonical measure for comparing probability distributions, enabling the definition of relative entropy

$$H_\rho[X] = -\mathbb{E}_\mu[\log \frac{\mathrm{d}\mu}{\mathrm{d}\rho}]. \tag{I.3}$$

Additionally, it provides a coordinate–independent way to specify symmetry transformations, ensuring that symmetry definitions remain consistent across different parameterizations of the same statistical manifold. This can be analogized to phase space shifting operations that leave the Gibbs integration measure invariant can be understood as gauge transformations, with the reference density playing the role of a gauge fixing condition.**[Phys. Rev. Lett. 133, 217101 (2024) –KD]**

The mathematical machinery for statistical symmetries centers on how probability densities transform under coordinate changes. Under a transformation $X = g(Y)$, probability densities transform as

$$p_y(y) = p_x(g(y))|\det(g'(y))| \tag{I.4}$$

This transformation law, involving the Jacobian determinant $|\det(g'(y))|$, ensures probability conservation. The Jacobian factor measures how volumes scale under transformation. $|\det(J)| = 1$ characterises volume preserving transformations like rotations and reflections.

For a probability distribution to exhibit symmetry under group action G, it must satisfy the condition

$$\forall g \in G \, \forall x \in X \, p(x) = p(g(x)) \, |\det g'(x)| \tag{I.5}$$

This condition is far more restrictive than point symmetry, as it demands global consistency across the entire probability measure.

The group-theoretic formulation provides additional structure. A group G acts on a probability space (Ω, F, μ) through measurable maps that preserve the $\sigma-$algebra structure.**[doi.org/10.1016/S1874-575X(06)80028-7 –KD]** Such a group can be decomposed using the orbit–stabilizer theorem, decomposing the action into orbits and stabilizers, and representation theory enables systematic construction of invariant functions and decomposition of function spaces into irreducible components.

I.C.3 Applications to particle physics

Particle physics provides a rich domain for applying statistical symmetries, particularly in detector calibration, data unfolding, and inference tasks at collider experiments.**[arXiv:2009.14613 –KD]** In detector response modeling, statistical symmetries constrain how identical particles must be treated. **[doi.org/10.17226/6045 –KD]** Permutation symmetry requires that detector analysis respect the indistinguishability of identical particles, affecting event reconstruction algorithms, particle identifi-

cation methods, and background estimation techniques. **[doi.org/10.1073/pnas.93.25.1425 –KD]** Response matrices used in unfolding procedures must respect particle exchange symmetries, with regularization procedures preserving these symmetry properties.

Unfolding represents an interesting application where statistical symmetries could guide the correction of detector effects. The process must preserve the statistical symmetries of the underlying physics, maintaining correlations between identical particles and ensuring conservation laws implied by symmetries remain valid. Full phase space unfolding could be facilitated by statistical symmetries that constrain unfolding procedures in high-dimensional phase spaces. Background estimation methods also exploit symmetry properties through carefully designed control regions, while systematic uncertainties account for potential symmetry violations.

In this way statistical symmetries fundamentally shape measurement and inference tasks through multiple pathways. Phase space shifting operations that preserve physically meaningful quantities have even been used outside of HEP to develop new computational approaches for molecular simulations.**[doi.org/10.1038/s42005-021-00669-2 –KD]** These symmetries lead to exact correlation relations between forces and observable properties, offering systematic ways to derive new statistical relationships. In machine learning too, encoding known symmetries dramatically improves performance. Symmetry encoding reduces sample complexity exponentially, with multidimensional symmetries providing disproportionately large

returns.**[2303.14269 –KD]** Equivariant neural networks, designed to respect known symmetries through architectural constraints, achieve superior data efficiency and generalization.**[2409.07327 –KD]**

The theoretical framework connects reference measures, symmetry principles, and statistical inference in profound ways. Reference measure selection can be guided by symmetry considerations, leading to principled approaches for prior selection in Bayesian inference. **[DOI:10.2307/1968511 –KD]** Symmetry principles suggest natural classes of invariant estimators, while statistical analogs of physical conservation laws emerge from symmetry considerations, constraining the behavior of statistical systems.**[10.1073/pnas.93.25.14256 –KD]** This unified framework offers concrete advantages for computational statistics and machine learning while opening new directions for both theoretical development and practical applications.

I.D SymmetryGAN: Discovering Symmetries with Adversarial Learning

The challenge of automatically discovering symmetries in high dimensional data represents a fundamental problem at the intersection of physics and machine learning. While physicists have long relied on theoretical insight to identify symmetries, the explosion of complex data from modern experiments demands automated approaches. SymmetryGAN emerges as an elegant solution, leveraging the power of adversarial learning to discover hidden invariances without prior knowledge of their existence.

The core insight driving SymmetryGAN is deceptively simple. If a transformation truly represents a symmetry of a dataset, then a well trained discriminator should not be able to distinguish between the original data and its transformed counterpart. This principle, when combined with the rigorous framework of inertial reference densities developed in the previous section, yields a powerful methodology for symmetry discovery that is both theoretically grounded and practically effective.

I.D.1 The SymmetryGAN Architecture

The SymmetryGAN framework modifies the traditional generative adversarial network architecture in a fundamental way. Rather than mapping from random noise to structured data, SymmetryGAN maps from data to data, learning transformations

that preserve the underlying probability distribution. Like a traditional GAN, however, SymmetryGAN consists of two neural networks engaged in adversarial training.

1. A generator $g : \mathbb{R}^n to \mathbb{R}^n$ that learns symmetry tranformations, and

2. A discriminator $d : \mathbb{R}^n to [0, 1]$ that attempts to distinguish original from transformed data.

The generator network g parametrizes potential symmetry transformations of the input space. Unlike traditional GANs where the generator creates new samples, here it transforms existing data points. The discriminator d plays its familiar adversarial role, attempting to classify whether a given sample comes from the original dataset or has been transformed by g.

The training objective is related but different from standard GANs. The loss function is

$$L[g, d] = -\frac{1}{N} \sum_{i=1}^{N} [\log(d(x_i)) + \log(1 - d(g(x_i)))] \tag{I.6}$$

This formulation differs from the usual binary cross entropy in that the same data samples appear in both terms, creating a self–adversarial structure. The generator seeks transformations that maximally confuse the discriminator, while the discriminator learns to detect deviations from the true data distribution.

By varying the loss with respect to the neural networks, one can show that at the theoretical optimum, the optimal discriminator is

$$d_*(x) = \frac{p(x)}{p(x) + p(g_*(x))|g'_*(x)|} = \frac{1}{2} \tag{I.7}$$

and the optimal generator obeys

$$p(x) = p(g_*(x))|g'_*(x)|. \tag{I.8}$$

i.e.the discriminator is maximally confused with equal probability of classifying any sample as original or transformed, and the generator obeys precisely the definition of a statistical symmetry.

The architecture naturally handles both discrete and continuous symmetries. For discrete symmetries like reflections or cyclic groups, the generator learns specific transformation matrices. For continuous symmetries like rotations, it parametrizes the corresponding Lie group elements. This flexibility makes SymmetryGAN applicable across diverse physical systems.

[fig:symmetrygan-architecture –KD] shows the modified GAN architecture of SymmetryGAN.

I.D.2 Machine Learning with Inertial Restrictions

The incorporation of inertial reference densities, as established in Section I.C, presents both theoretical necessity and practical challenges. These restrictions can be implemented through three distinct approaches.

1. Simultaneous Discrimination: In this approach, discriminators evaluate transformations on both the target dataset and samples from the inertial density. The loss function is extended to include terms from the inertial distribution p_I

$$L_{\text{sim}}[g, d, d_I] = L[g, d] + \lambda L_I[g, d_I] \tag{I.9}$$

   where d_I is a separate discriminator network specialized for the inertial distribution. This method offers maximal flexibility, allowing discovery of symmetries even when the inertial distribution's symmetries aren't known analytically. However, it requires the ability to sample from p_I which becomes problematic for improper priors like uniform distributions on $\mathbb{R}^n$.

2. Two–Stage Selection: This approach involves first identifying all PDF–preserving maps, then filtering for those preserving the inertial density. The two–stage approach decouples the symmetry discovery problem into training multiple generators to find PDF–preserving transformations of the target data and *post hoc* verifying which transformations also preserve p_I While conceptually clean,

this method proves computationally wasteful, as the space of PDF–preserving maps vastly exceeds the space of true symmetries.

3. Upfront Restriction: Here one constrains the generator architecture to only produce transformations that preserve the inertial density by construction in the first place, using prior knowledge about p_I.

It is this third approach that is adopted in the SymmetryGAN implementation. When the inertial distribution is uniform on $\mathbb{R}^n$ the generator is restricted to equiareal maps—transformations with unit Jacobian determinant. For affine transformations, this corresponds to the affine special linear group $\mathbb{A}SL_n^{\pm}(\mathbb{R})$ and its extensions.

The upfront restriction method offers computational efficiency gains by searching only among valid symmetrie in addition to theoretical guarantees that discovered transformations are true symmetries. It also simplifies training without multiple discriminators or *post hoc* verification.

The restriction to linear equiareal maps, while limiting, captures a rich class of symmetries including rotations, reflections, shears, and all their compositions. The Iwasawa decomposition**[–KD]** provides a systematic parametrization of $g \in \mathbb{A}GL_n(\mathbb{R})$.

$$g = \sqrt{|d|} I^{\frac{1-\mathrm{sgn}(d)}{2}} \cdot R(\theta) \cdot D(r) \cdot S(u) + v \tag{I.10}$$

where d is the determinant, I is the involution matrix, $R(\theta) \in SO_n(\mathbb{R})$ is a rotation, $D(r)$ is a dialatation, $S(u)$ is a sheer, and $v \in \mathbb{R}^n$ is a translation. As an illustration, in two dimensions,

$$I = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \quad R(\theta) = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \quad D(r) = \begin{bmatrix} r & 0 \\ 0 & \frac{1}{r} \end{bmatrix} \quad S(u) = \begin{bmatrix} 1 & u \\ 0 & 1 \end{bmatrix} \quad v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \tag{I.11}$$

This decomposition enables efficient exploration of the symmetry space through gradient descent.

I.D.3 Deep learning implementation details.

The practical implementation of SymmetryGAN requires careful attention to architectural choices, training dynamics, and numerical stability. The framework's success depends on balancing the adversarial training process while maintaining the theoretical properties that enable symmetry discovery.

For linear symmetries, the generator directly parametrizes transformation matrices. Rather than using fully connected layers, the generator outputs parameters of the Iwasawa decomposition, ensuring all generated transformations lie within the constrained search space. This architectural choice provides interpretability, where each parameter has clear geometric meaning, stability, by avoiding ill-conditioned matrices

through structured parametrization, and efficiency, by reducing the parameter count compared to unconstrained matrices

For discovering specific symmetry subgroups, the loss function can further incorporate additional constraints. For example, to find cyclic symmetries $\mathbb{Z}_q$ the loss can be augmented to encourages $g^q = \mathrm{id}$.

$$L_{\text{cyclic}}[g, d] = L_{\text{BCE}}[g, d] + \alpha \sum_i ||g^q(x_i) - x_i||^2. \tag{I.12}$$

The discriminator architecture follows standard practices for the data domain. For low–dimensional examples, simple feedforward networks with $2 - 3$ hidden layers suffice. †he discriminator need not be overly complex. Its role is to provide gradient signal for the generator's symmetry search, not to achieve perfect classification.

The optimization landscape analysis reveals its topological structure. For simple distributions like Gaussians, the loss landscape contains distinct maxima corresponding to each symmetry, separated by deep valleys. This topology explains why random initialization reliably discovers different symmetries. The generator converges to the nearest local maximum, which corresponds to a true symmetry.

The framework demonstrates remarkable robustness to hyperparameter choices. Across experiments with Gaussian mixtures and particle physics data, a range of hyperparameter settings achieve consistent symmetry discovery.

I.D.4 Verification Protocols

A few different checks can be applied to verify that the discovered transformation represents a true symmetry.

- Visual inspection: Visually inspecting low dimensional slices of the original and transformed data serves as a simple sniff test.
- Loss value: True symmetries always achieve a loss of $2 \log 2$ at convergence.
- Distribution matching: Statistical divergences between X and $g(X)$ such as the KL divergence measure whether the two datasets have the same probability density or not.
- Invarian verification: known invariants, such as moments, should be identical between the distributions.

I.D.5 Other Symmetry Discovery Methods

The landscape of symmetry discovery in machine learning has evolved considerably, with approaches ranging from classical statistical methods to modern deep learning architectures. Understanding SymmetryGAN's position within this ecosystem illuminates both its unique contributions and its connections to broader methodological trends.

Classical statistical approaches traditionally relied on hypothesis testing and moment matching. These methods typically test specific, pre–defined symmetry hypotheses by relying on low-dimensional projections or summary statistics. They struggle with high-dimensional data or complex symmetry groups and require extensive domain knowledge to formulate appropriate tests.

The Cramer-von Mises and Anderson-Darling tests**[cite –KD]** exemplify this approach, checking whether transformed data follows the same distribution as the original. While rigorous, these methods scale poorly and cannot discover unexpected symmetries.

Early machine learning methods introduced automation to the symmetry discovery process. Methods like Group-Invariant Autoencoders**[cite –KD]** learn representations invariant to known symmetry groups but cannot discover unknown symmetries Spatial Transformer Networks**[cite –KD]** learn task-specific transformations but don't explicitly identify symmetries *per se*. Maximum Mean Discrepancy approaches**[cite –KD]** and other kernel methods can test symmetry but struggle with continuous groups.

Contemporary deep learning methods have introduced a range of innovations to the field. Lie Group Learning Networks directly parametrize Lie algebra generators**[cite –KD]**. Such methods have excellent theoretical grounding, and physically interpretable parameters. They are, however, restricted to continuous symmetries, and require

complex optimization for convergence. Latent LieGAN (LaLiGAN) pioneered the discovery nonlinear symmetries via latent space linearization**[cite –KD]**. It enabled the field to make strides in handling nonlinear group actions, but the additional complexity leads to potential latent space artifacts Reinforcement Learning enables automatic augmentation discovery through policy learning**[cite –KD]**. The method is flexible in being task–performance driven, and it handles discrete symmetries well.

SymmetryGAN exemplifies several important trends in modern machine learning for HEP. Methods have broadly benefited from incorporating domain knowledge through architectural constraints and loss terms. In general the rise of equivariant eearning has driven the move beyond invariance to discover the underlying symmetry structure. SymmetryGAN also exemplies the trend towards interpretability by producing human-understandable symmetry parameters rather than opaque features. The method is however limited in its restriction to parametrizable symmetry classes. It also exclusively enables symmetry discovery, the task of identifying elements of the symmetry group of a dataset. Although the paper does provide some steps towards symmetry inference, the task of identifying and classifying the symmetry group as a whole, it remains challenging to infer complete symmetry group structure from discovered elements. The field continues to evolve rapidly, with SymmetryGAN representing a significant step toward automated, theoretically grounded symmetry discovery. Its success in particle physics applications, detailed in the following section,

demonstrates the practical value of this principled approach to uncovering hidden invariances in complex data.

I.E Empirical Experiments

The theoretical machinery of SymmetryGAN, with its rigorous statistical framework and inertial reference densities, proves its worth through empirical validation. Moving from abstract formulations to concrete applications reveals both the method's surprising effectiveness and the subtle challenges that emerge with real data. This section details the application of this method on carefully controlled Gaussian experiments to complex collider data, demonstrating how SymmetryGAN bridges the gap between mathematical elegance and practical discovery.

I.E.1 Gaussian Experiments

The choice to begin with Gaussian distributions reflects more than mere mathematical convenience; it represents a strategic approach to validating a novel methodology. Gaussians offer analytically tractable loss landscapes that allow direct comparison between theoretical predictions and empirical results, serving as a crucial proof of concept before tackling high dimensional HEP data.

I.E.1.i One-Dimensional Gaussian

The simplest non–trivial test case consists of a Gaussian distribution $\mathcal{N}(0.5, 1.0)$ with a reflection symmetry about $x = 0.5$. This apparently simple example harbors

surprising richness. The disconnected nature of the symmetry space in even this simple example foreshadows the challenges of symmetry inference. Discovering individual symmetries differs fundamentally from understanding their group structure.

The distribution admits precisely two linear symmetries: the identity transformation $g(x) = x$ and the reflection $g(x) = 1 - x$. When the generator is parametrized as $g(x) = b + cx$, the analytic loss landscape has the topology shown in **[fig:Z2analytic –KD]**. The two maxima corresponding to these symmetries are separated by a deep valley at $c = 0$, creating a disconnected solution space. The empirical results validate the theoretical predictions with remarkable precision. Random initialization of parameters $(b_i, c_i) \sim \mathcal{U}([-5, 5]^2)$ leads to convergence at one of two distinct clusters: $(b_f, c_f) = (0, 1)$ for the identity or $(1, -1)$ for the reflection as can be seen in **[fig:Z2numeric –KD]**. The loss barrier at $c = 0$ acts as a watershed, deterministically routing the optimization based on the sign of the initial slope. This behavior illuminates a fundamental principle: the loss landscape topology directly determines the discoverable symmetry structure.

I.E.1.ii Two-Dimensional Gaussians

The two-dimensional Gaussian examples increase the complexity while maintaining analytical tractability. Consider first the standard bivariate normal $N_{1,1} = \mathcal{N}(\mathbf{0}, \nVdash_2)$, which possesses the full orthogonal group $O(2)$ as its symmetry group. When the

generator is restricted to the form

$$g(x) = \begin{bmatrix} c & s \\ -s & c \end{bmatrix} x \tag{I.13}$$

SymmetryGAN must discover that only transformations satisfying $c^2 + s^2 = 1$ represent true symmetries as can be seen in **[fig:SO2-i –KD]**. The empirical results demonstrate the method's ability to learn this constraint without explicit enforcement. Starting from random initializations within $[-1, 1]^2$, the learned parameters consistently converge to the unit circle, validating that SymmetryGAN can discover not just discrete symmetries but continuous symmetry manifolds**[cite –KD]**. This represents a significant achievement: the neural network learns the algebraic constraint defining $SO(2)$ purely from data.

The two dimensional Gaussian example also allows us to test the approach described in Eq. (I.12). **[fig:cyclic-loss –KD]** shows the discovered symmetries for the $N_{1,1}$ distribution when the loss function is augmented with the constraint Eq. (I.12), for $q = 2, 3$, and 7, with $\alpha = 0.1$. Both the analystic loss landscape and SymmetryGAN's output confirm that the continuous $SO(2)$ symmetry is split into discrete symmetries at the q^{th} roots of unity upon the inclusion of the mean squared error term.

One could also consider a two dimensional Gaussian with a non–trivial covariance matrix, such as $N_{1,2} = \mathcal{N}(mathbf0, \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix})$. The full symmetry group of this distribution is highly non–trivial and is described below, but among other features, it contains as a subgroup the Klein 4−group $V_4 = \{⊮, -⊮, \sigma_3, -\sigma_3\}$ for the Pauli matrix σ_3. When SymmetryGAN is limited to transformations of the form **??**, it discovers the Klein 4−group V_4 as demonstrated in **[fig:V4 –KD]**. When the generator is allowed to search for general linear transformations inside $GL_2(\mathbb{R})$, it discovers the full symmetry group that comprises the afforementioned Klein 4−group as well as transformations involving dilatations.

[fig:agl2 –KD] shows the discovered symmetries for the $N_{1,1}$ distribution and **[fig:agl2-2 –KD]** shows the discovered symmetries for $N_{1,2}$ when the generator is able to discover general linear transformations inside $\mathbb{A}GL_2(\mathbb{R})$.

I.E.1.iii Gaussian Mixture Models

The power of SymmetryGAN becomes more apparent when applied to Gaussian mixture models with complex symmetry groups. Three such examples are considered, inspired by the the examples in kdcite. The first is a one dimensional bimodal distribution,

$$p(x) = \frac{1}{2}\mathcal{N}(-1, 1) + \frac{1}{2}\mathcal{N}(1, 1) \tag{I.14}$$

which possesses the cyclic group $\mathbb{Z}_2 = \{x \mapsto \pm x\}$ as its symmetry group. The discovered symmetries are shown in **[fig:otherdistributions-1D –KD]**.

The second is a two dimensional octagonal distribution,

$$p(x) = \frac{1}{8}\sum_{i=1}^{8} \mathcal{N}(\cos\frac{2\pi i}{8}, 0.1) \times \mathcal{N}(\sin\frac{2\pi i}{8}, 0.1) \tag{I.15}$$

which has the dihedral group D_8 as its symmetry group, a non-commutative discrete group containing both rotations and reflections. The third is a 5×5 square distribution,

$$p(x) = \frac{1}{25}\sum_{i=0}^{4}\sum_{j=0}^{4} \mathcal{N}(i-2, 0.1) \times \mathcal{N}(j-2, 0.1) \tag{I.16}$$

which has the dihedral group D_4 as its symmetry group, also a non-commutative discrete group containing both rotations and reflections. The left column of **[fig:otherdistributions-2D –KD]** shows data from each of these distributions, the middle column shows the discovered rotations, and the right column shows the discovered reflections. In each case SymmetryGAN is able to correctly discover the symmetries of the distribution.

I.E.2 LHC Dijet Experiments

The transition from Gaussian toy models to LHC data represents a quantum leap in complexity. Where the Gaussian examples involved simple, low dimensional parameter spaces that could be analyzed analytically, dijet events exist in a high dimensional phase space, with each event containing variable numbers of particles

subject to complex kinematic constraints**[cite –KD]**. For this example, we will study the transverse momenta of dijet events from CMS Open Data.**[cite –KD]** Each event is characterised by the transverse momentum components

$$\mathbf{x} = (p_{1x}, p_{1y}, p_{2x}, p_{2y}) \tag{I.17}$$

The initial exploration considers transformations from $SO(2) \times SO(2)$, independently rotating each jet in the transverse plane. The generator is restricted to the form

$$g_{\theta_1,\theta_2}(\mathbf{x}) = \begin{bmatrix} R(\theta_1) & 0 \\ 0 & R(\theta_2) \end{bmatrix} \mathbf{x} \tag{I.18}$$

Momentum conservation predicts that only simultaneous rotations $\theta_1 = \theta_2$ should preserve the distribution. The empirical results confirm this beautifully: starting from random initializations in $[0, 2\pi)^2$, the learned parameters cluster along the diagonal $\theta_1 = \theta_2$. **[fig:LHCO-i –KD]** shows the discovered symmetries for the $SO(2) \times SO(2)$ case. The symmetry discovery map for this system reveals unexpected elegance. The mapping from initial to final parameters follows:

$$\Omega(\theta_1, \theta_2) = \begin{cases} \frac{\theta_1+\theta_2}{2} & |\theta_1 - \theta_2| < \pi \\ \frac{\theta_1+\theta_2}{2} - \pi & |\theta_1 - \theta_2| > \pi \end{cases} \tag{I.19}$$

This bisection of the angular difference represents the path of steepest ascent in the loss landscape, demonstrating how gradient dynamics naturally discover the constraint imposed by momentum conservation**[fig:LHCO-ii –KD]**.

Expanding to the full $SO(4)$ search space introduces greater complexity. An arbitrary element of $SO(4)$ is the composition of six independent rotations, in the six possible orthogonal $2-$planes. The six-parameter group admits no simple visualization, and the discovered symmetries no longer lie in any two– or three–dimensional subspace. Validation requires more sophisticated statistical tests. Visual inspection can serve as a smell test. Projections of the original and transformed distributions onto physically meaningful observables (jet p_T, azimuthal angles) show excellent agreement**[fig:LHCO-Comparison –KD]**. Having passed the visual test, we can now apply more discerning tests to ensure the symmetries are not spurious. The transformed azimuthal angles

$$\begin{bmatrix} \tilde{\phi}_1 \\ \tilde{\phi}_2 \end{bmatrix} = \begin{bmatrix} \arctan \frac{g(\theta_1,\theta_2)(p_{1y})}{g(\theta_1,\theta_2)(p_{1x})} \\ \arctan \frac{g(\theta_1,\theta_2)(p_{2y})}{g(\theta_1,\theta_2)(p_{2x})} \end{bmatrix} \tag{I.20}$$

are uniformly distributed only for true symmetries, and not for any other transformation, as shown in **[fig:phis –KD]**. Computing the KL divergence between the transformed and original distributions is another way to test for symmetry. The KL divergence should be zero for true symmetries, and non-zero for any other transformation. While this is neveer exactly true, **[fig:KL-symm –KD]** shows the KL divergence for true symmetries approach the irreducible statistical noise floor. Finally, *post hoc* discriminators trained on discovered transformations achieve losses within 1% of the theoretical optimum $2\log 2$ confirming these represent genuine

symmetries**[fig:LHCOLosses –KD]**. The discovered subgroup of $SO(4)$ includes the expected $SO(2)$ rotations reflecting cylindrical detector geometry, but also unexpected combinations that would be difficult to identify through physical intuition alone.

I.E.3 Interpreting Discovered Symmetries

The interpretation of discovered symmetries requires bridging the gap between abstract mathematical transformations and physical understanding. This translation process reveals both the power and limitations of automated symmetry discovery.

Each discovered symmetry transformation corresponds to a conservation law or invariance principle in the underlying physics. For the dijet system, the discovered azimuthal rotations connect directly to angular momentum conservation about the beam axis. The continuous $SO(2)$ symmetry reflects the absence of any preferred direction in the transverse plane—a fundamental consequence of the cylindrical geometry of both the collision process and the detector**[cite –KD]**.

More subtle are the discovered discrete symmetries. The non$-SO(2)$ transformations found combine specific overall rotations with jet exchanges. For example,

$$g_{Z_2}(p_{1x}, p_{1y}, p_{2x}, p_{2y}) = (-p_{2x}, -p_{2y}, -p_{1x}, -p_{1y}) \tag{I.21}$$

is a transformation that simultaneously rotates by π and exchanges the two jets—a symmetry that follows from the identical nature of QCD jets but would be non–obvious

without systematic search. The relationship between discovered symmetries and known physics provides crucial validation. Each symmetry found by SymmetryGAN corresponds to either:

- A fundamental symmetry of physics (Lorentz invariance, gauge symmetry)
- A symmetry of the experimental setup (cylindrical detector geometry)
- An emergent symmetry from kinematic constraints (momentum conservation)

No spurious symmetries appear, demonstrating the method's reliability. However, the challenge remains: how do we move from discovering individual symmetry transformations to understanding the complete symmetry group structure?

I.E.4 Towards Symmetry Inference

The distinction between symmetry discovery and symmetry inference represents a fundamental challenge in the SymmetryGAN framework. While the method excels at finding individual elements of a symmetry group, reconstructing the complete group structure requires additional theoretical machinery**[cite –KD]**.

The task of symmetry inference can be formulated as "Given a set of discovered symmetry transformations $\{g_1, g_2, ..., g_n\}$, determine the minimal group G containing

all these elements." Three complementary approaches emerge for tackling this challenge.

1. Discrete Subgroup Analysis: By modifying the loss function to enforce specific algebraic relations, SymmetryGAN can systematically probe for discrete subgroups. The cyclic group constraint:

$$L_{\text{cyclic}}[g, d] = L_{\text{BCE}}[g, d] + \alpha \sum_i ||g^q(x_i) - x_i||^2 \tag{I.22}$$

successfully identifies Z_q subgroups within larger symmetry groups**[fig:MSE –KD]**. This approach extends naturally to other discrete groups through appropriate constraint terms, though non-Abelian groups require more sophisticated loss modifications**[cite –KD]**.

2. Group Composition Methods: The closure property of groups suggests an iterative approach: given discovered symmetries $\{g_i\}$, compute all possible compositions $g_i \circ g_j$ to expand the known group elements. For continuous groups like $SO(2)$, a single irrational rotation angle generates a dense subgroup, effectively recovering the entire group through composition**[cite –KD]**. However, numerical precision limits this approach. Each composition compounds errors, leading to degradation after multiple iterations. The practical limit appears to be 10-20 compositions before accumulated errors overwhelm the signal.

3. The Symmetry Discovery Map: Perhaps the most promising direction involves learning the complete mapping from initialization space to symmetry space:

$$\Omega : \mathbb{R}^k \to \mathcal{G} \tag{I.23}$$

where $\mathcal{G}$ represents the symmetry group manifold. This map encodes the complete symmetry structure implicitly, as its image is precisely the symmetry group**[cite –KD]**. Preliminary attempts to learn Ω had limit success. The main challenges are that the map must begin near the identity to preserve the connection between initial and final parameters, and the min–max nature of GAN training becomes more complex when optimizing over transformation maps rather than transformations themselves. Despite these challenges, successful learning of even approximate symmetry discovery maps would provide us with a powerful tool to characterize symmetry groups.

I.F Symmetry Informed Unfolding

The fundamental challenge of unfolding, recovering truth–level distributions from detector–smeared observations, stems from its ill–posed nature. Multiple truth distributions can yield identical observations after detector effects, creating an underdetermined system crying out for additional constraints. Symmetry provide a natural regularization scheme for unfolding problems. When a physical process respects a symmetry, that invariance constrains the space of valid solutions, effectively regularizing the inverse problem without introducing artificial biases. Consider the standard unfolding problem formulated as an optimization,

$$\min_{p_{\text{truth}}} \left[\chi^2(p_{\text{obs}}, R \cdot p_{\text{truth}}) + \lambda \cdot \text{Reg}(p_{\text{truth}})\right] \tag{I.24}$$

where R represents the response matrix and Reg is a regularization term. Traditional approaches use generic smoothness penalties, but symmetries provide physically motivated constraints. For a symmetry group G discovered through SymmetryGAN, one could construct the symmetry–preserving regularization,

$$\text{Reg}_{\text{sym}}(p) = \sum_{g \in G} \|p - g \cdot p\|^2 \tag{I.25}$$

This penalty vanishes for distributions respecting the symmetry while penalizing asymmetric solutions. Such a regularization term would enforce invariances that the

underlying physics demands**[cite –KD]** rather than imposing arbitrary smoothness constraints.

However such an implementation would have to address subtle challenges. For continuous symmetries like $SO(2)$, the sum over group elements becomes an integral requiring careful handling. A possible solution is to select group elements that uniformly cover the symmetry manifold. For $SO(2)$, for instance, one could sample angles $\theta_i = 2\pi i/N$ for $i = 0, ..., N-1$. However, any such choice would be *ad hoc* and woudl introduce artifacts into the unfolding process.

For non-compact groups, the challege is further compounded by the fact that there no longer exists a equivariant sampling strategy in the first place. Methods such as importance sampling based on data distribution have been proposed**[cite –KD]**, but they remain to be tested in the physics context.

The theoretical foundation of this approach connects to information theory. Each symmetry represents a constraint reducing the effective dimensionality of the solution space. For a distribution over N bins with a $\mathbb{Z}_k$ cyclic symmetry, the degrees of freedom reduce from N to N/k, improving the effective statistics by a factor of k. This dimensionality reduction should translate directly to improved unfolding stability**[cite –KD]**.

I.F.1 Symmetry Preserving Neural Network Architectures

The architectural encoding of symmetries represents a paradigm shift from *post hoc* constraints to intrinsic guarantees. A neural network f_θ is said to be equivariant to group G if for all $g \in G$ and inputs x

$$f_\theta(g \cdot x) = g \cdot f_\theta(x) \tag{I.26}$$

For unfolding applications, we could seek architectures where the mapping from observed to truth–level distributions preserves known symmetries. The Group Convolutional Neural Network (G-CNN) framework provides one example of such an approach**[cite –KD]**. For a symmetry group G acting on the input space, G-CNN layers take the form

$$f * \psi = \int_H f(h)\psi(g^{-1}h)dh \tag{I.27}$$

where $*$ denotes group convolution and H is the stabilizer subgroup. This formulation guarantees equivariance by construction, not training.

The LGN (Lorentz Group Network) architecture**[cite –KD]** constructs features from Lorentz scalars and 4-vectors in order to preserve Lorentz invariance. For identical particle handling, Deep Sets**[cite –KD]** and Set2Graph**[cite –KD]** provide principled approaches in which

$$f(\{x_1, ..., x_n\}) = \rho\left(\sum_i \phi(x_i)\right) \tag{I.28}$$

where ϕ and ρ are learned functions, guaranteeing invariance to particle ordering. For processes involving gauge bosons, architectures must respect gauge transformations. The Gauge Equivariant Mesh CNN**[cite –KD]** provides a framework for $SU(3) \times SU(2) \times U(1)$ invariance.

A hybrid architecture for symmetry-aware unfolding could combine these elements. The architecture would necessarily ensure that symmetries present in the observed data propagate to the unfolded distribution. However, the practical implementation of such an architecture remains an open question.

I.F.2 Reducing Dimensionality Through Symmetry Identification

The curse of dimensionality poses a significant challenge to unfolding problems. Each additional degree of freedom exponentially increases the solution space, degrading statistical power and amplifying instabilities. Symmetry identification offers a principled method for revealing redundancies that can be eliminated without information loss.

For a distribution with symmetry group G, the effective dimensionality reduces from the ambient space dimension to the dimension of the quotient space X/G. A function $f : X \to \mathbb{R}$ is G-invariant if and only if there exists a function $\tilde{f} : X/G \to \mathbb{R}$

with $f = \tilde{f} \circ \pi$ where $\pi : X \to X/G$ is the canonical projection. This factorization is the key to dimensionality reduction. One needs only unfold $\tilde{f}$ on the lower–dimensional quotient space.

Any such process of symmetry-based projection, however, requires careful treatment of statistical uncertainties. The optimal projection minimizes information loss while maximizing dimensionality reduction. The statistical gain from this procedure scales dramatically with dimensionality. For a d-dimensional space with a k-parameter continuous symmetry group, the effective dimensionality reduces to $d - k$, yielding an effective statistics increase of

$$N_{\text{eff}} = N \times \left(\frac{V_{\text{full}}}{V_{\text{reduced}}} \right)^{1/d} \tag{I.29}$$

where V denotes phase space volumes**[cite –KD]**.

I.F.3 Hidden Symmetries and Emergent Simplicity

Often, the most powerful dimensionality reductions come from unexpected symmetries. The discovery of approximate custodial symmetry in electroweak physics simplified phenomenology dramatically. Similarly, emergent symmetries in complex final states can enable radical dimensionality reduction**[cite –KD]**. The quantitative impact of symmetry on unfolding precision admits rigorous mathematical treatment. By connecting group theory, information theory, and statistical estimation,

we can derive fundamental bounds on the achievable precision gains from symmetry constraints.

For an unbiased estimator $\hat{\theta}$ of parameters θ with symmetry group G, the covariance satisfies

$$\text{Cov}(\hat{\theta}) \geq \frac{1}{N}\mathcal{I}_G^{-1}(\theta) \tag{I.30}$$

where $\mathcal{I}_G$ is the $G-$equivariant Fisher information matrix.**[cite –KD]**

The symmetry constraints modify the Fisher information by projecting onto the $G-$invariant subspace. For a likelihood $L(\theta|x)$ with symmetry group G, the equivariant Fisher information takes the form

$$[\mathcal{I}_G]_{ij} = \mathbb{E}\left[\frac{\partial \log L}{\partial \theta_i}\frac{\partial \log L}{\partial \theta_j}\right] - \sum_{g\in G} w_g \mathbb{E}\left[\frac{\partial \log L}{\partial \theta_i}\frac{\partial \log L(g\cdot\theta)}{\partial \theta_j}\right] \tag{I.31}$$

where w_g are group averaging weights**[cite –KD]**.

This modified Fisher information leads to a tighter Cramér-Rao Bound on parameter estimation.**[cite –KD]** For the specific case of unfolding with a discrete symmetry group G of order $|G|$, the variance reduction factor would be

$$\frac{\text{Var}(\hat{\theta}_{\text{sym}})}{\text{Var}(\hat{\theta}_{\text{unconstrained}})} = \frac{1}{|G|} + \frac{|G|-1}{|G|}\rho \tag{I.32}$$

where ρ is the average correlation between symmetric images of the parameter**[cite –KD]**.

I.F.4 Application to Jet Substructure Unfolding

Consider unfolding jet substructure observables like N-subjettiness ratios τ_{21}. The IRC (infrared-collinear) safety of these observables implies approximate scale invariance. For a scaling symmetry with parameter λ:

$$\tau_{21}(\lambda p_T, \lambda m_{\text{jet}}) = \tau_{21}(p_T, m_{\text{jet}}) \tag{I.33}$$

This symmetry reduces the 2D unfolding problem in (p_T, m_{jet}) to a 1D problem in the ratio m_{jet}/p_T. The maximal possible theoretical precision gain is

$$\sigma_{\tau_{21}}^{\text{sym}} = \frac{\sigma_{\tau_{21}}^{\text{2D}}}{\sqrt{\log(p_T^{\text{max}}/p_T^{\text{min}})}}. \tag{I.34}$$

For typical jet p_T ranges spanning two orders of magnitude, this yields a $4.6 \approx 2.1\times$ improvement in precision**[cite –KD]**.

Such an information–theoretic analysis also provides additional insights. Each symmetry can theoretically reduce the Kullback-Leibler divergence between truth and reconstructed distributions by as much as

$$\Delta D_{KL} = \frac{1}{2}\log|G| - \frac{1}{2N}\text{Tr}(\mathcal{I}_G^{-1}\mathcal{I}_{\text{unconstrained}}) \tag{I.35}$$

This reduction could directly translate to improved unfolding fidelity, with larger symmetry groups yielding greater gains**[cite –KD]**.

I.F.5 Case Study: Improving NPU with Symmetry Constraints

The Neural Posterior Unfolding (NPU) framework**[cite –KD]**, discussed in Chapter 7, provides an interesting potential testbed for symmetry integration. By incorporating discovered symmetries into NPU's generative model, it might be possible to achieve substantial improvements in both computational efficiency and physical fidelity. One could modify the NPU generative model $g_\phi(z, x_{\text{obs}})$ to respect discovered symmetries through architectural constraints and augmented training. The standard NPU loss function,

$$\mathcal{L}_{\text{NPU}} = \mathbb{E}_{p(x_{\text{obs}}, x_{\text{truth}})} \left[\log q_\phi(x_{\text{truth}} | x_{\text{obs}}) \right], \tag{I.36}$$

would become, with symmetry constraints,

$$\mathcal{L}_{\text{Sym-NPU}} = \mathcal{L}_{\text{NPU}} + \lambda \sum_{g \in G} \left\| g_\phi(z, x_{\text{obs}}) - g \cdot g_\phi(g^{-1} \cdot z, g^{-1} \cdot x_{\text{obs}}) \right\|^2 . \tag{I.37}$$

Such an augmented loss would drive the generative model to learn $G-$equivariant structures.

In effect the combination of NPU with SymmetryGAN has the potential to create a virtuous cycle, where better symmetry identification improves unfolding, and better unfolding reveals clearer symmetry patterns in the truth–level distributions.

Such a thought experiment points toward a broader principle. The most powerful machine learning approaches for physics don't just learn from data but incorporate fundamental physical principles into their architecture. By discovering and enforcing symmetries, we transform unfolding from a purely statistical exercise into a physics-informed inference problem, achieving results that respect both the data and the underlying theoretical framework.

I.G Symmetry Aware Unfolding for Improved Measurement Precision

The transformation of symmetry from abstract mathematical concept to concrete measurement tool represents one of the most profound opportunities in modern particle physics analysis. This section ventures into the frontier of how the marriage of symmetry awareness and unfolding might revolutionize our ability to extract truth from data. These ideas, while not yet implemented, chart a course toward a future where every discovered invariance translates directly into measurement precision.

I.G.1 Data Augmentation Using Discovered Symmetries

The concept of data augmentation through symmetry sits at a fascinating intersection of theoretical physics and practical statistics. The flip side of the dimensioanlity reduction discussed in Section I.F is that when we discover that a dataset respects certain transformations, we gain the ability to multiply our effective statistics without collecting a single additional event. This isn't merely a computational trick,it reflects a deep truth about the redundancy inherent in symmetric systems.

Symmetry based data augmentation is the process of generating statistically equivalent synthetic data by applying discovered symmetry transformations to existing events, effectively increasing sample size while preserving all physical properties.

Traditional data augmentation in machine learning often relies on heuristic transformations. These modifications hope to capture invariances without formal guarantees. By contrast, symmetry based augmentation leverages proven invariances discovered through methods like SymmetryGAN. Every augmented event is, by construction, as physically valid as the original.

As with the dimension reduction discussed in Section I.F, the subtlety lies in the sampling strategy. For finite groups, we might include all group elements, and for compact continous groups so long as we sample carefully to avoid bias from the uniform measure on the group manifold, the artifacts introducted will, in the very least, be unbiased. For non-compact groups, the uniform measure is not well defined, and there is no way to guarantee that the artifacts introduced are unbiased.

For a concrete example, consider unfolding the Z boson p_T spectrum with discovered azimuthal symmetry. Each measured event (p_T, ϕ, η) can generate a family of equivalent events $(p_T, \phi + \Delta\phi, \eta)$ for any $\Delta\phi$.

The power of symmetry based augmentation extends beyond simple counting statistics. Symmetry augmentation can heal detector defects. Imagine a calorimeter module at $\phi = \pi/2$ suffering from reduced efficiency. With symmetry augmentation, we can rotate events from healthy detector regions to fill the gap:

The corrected distribution is given by

$$n_{\text{corrected}}(\phi = \pi/2) = \frac{1}{|G|} \sum_{g \in G} \epsilon(g^{-1} \cdot \phi) \cdot n_{\text{observed}}(g^{-1} \cdot \phi) \tag{I.38}$$

where $\epsilon(\phi)$ represents the $\phi-$dependent efficiency. This is not merely interpolation, it's using physical symmetry to transport information across phase space.

The interplay with machine learning architectures reveals another dimension. Neural networks trained on symmetry–augmented data should exhibit

- Faster convergence: The effective sample size increase should translate directly to reduced training epochs.
- Better generalization: The network sees the full symmetry orbit during training, preventing overfitting to specific configurations.
- Automatic equivariance: Even non-equivariant architectures should learn approximate equivariance from augmented data.

For all the benefits that symmetry aware augmentation might provide, not all symmetries admit straightforward augmentation. Even with discrete symmetries like parity while we can flip event coordinates, care is needed with derived quantities. Or for instance, the jet clustering algorithm might produce different multiplicities when applied to parity-flipped events, breaking the supposed equivalence. In this sense the augmentation must respect the full analysis chain, not just the raw kinematics, and

symmetries that admit all the steps of the analysis chain are, naturally, more difficult to find.

I.G.2 Symmetry Constrained Unfolding

A variant of moment unfolding where the functional basis is restricted to symmetry respecting functions could guide choices of basis functions grounded in physical principles. In **??**, we discussed the p_T dependent Boltzmann–like basis functions, $\boldsymbol{\beta}_a(p_T)$, which were approximated as linear functions of jet p_T,

$$\boldsymbol{\beta}_a(p_T) = \beta_a^{(0)} + \beta_a^{(1)}\, p_T, \tag{I.39}$$

to constrain the flexibility of the generator network. A symmetry aware parametrization of the $\boldsymbol{\beta}_a(p_T)$ could provide a more principled way to constrain the generator network, while still achieving the necessary dimensionality reduction.

Implementing a similar approach in RAN architectures would requires careful attention to the increased complexity of the architecture, but a symmetry aware regularization scheme would provide improved conditioning, as the response matrix becomes block-diagonal in symmetry sectors, and physical consistency where unfolded distributions automatically respect, for example, IRC safety, providing smoothness within symmetry sectors without violating invariance.

The quantitative impact of symmetry awareness on measurement precision admits rigorous analysis through the lens of information theory and statistical estimation. Moving beyond hand-waving arguments about "effective statistics," we can derive precise bounds on the achievable improvements. The information content of data about parameters θ under symmetry group G is captured by the equivariant Fisher information matrix $\mathcal{I}_G(\theta)$, which bounds the achievable precision of any unbiased estimator. The fundamental result stems from the Equivariant Cramér-Rao Bound discussed in Section I.F.

In addition, symmetry awareness provides more than variance reduction. It fundamentally alters the bias-variance tradeoff. Traditional unfolding must balance bias from regularization (smoothing, early stopping) and variance from statistical fluctuations. Symmetry constraints modify the bias-variance tradeoff,

$$\text{Bias}_{\text{sym}} = \frac{\text{Var}_0}{n \cdot S(\mathcal{G})} \tag{I.40}$$

where $S(\mathcal{G})$ is the "symmetry factor" quantifying variance reduction, so that the total error becomes

$$\text{MSE}_{\text{total}} = \text{Bias}^2_{\text{sym}} + \frac{\text{Var}_0}{n \cdot S(\mathcal{G})}. \tag{I.41}$$

In this sense, symmetry aware unfolding isn't just about improved statistics—it's about extracting the maximum physical information from finite data.

I.H Conclusion

We began this chapter with the seemingly simple question of what it means for a dataset to possess symmetry, only to discover layers of mathematical subtlety involving Jacobian factors, inertial reference densities, and the delicate interplay between statistical and physical invariances. This final section looks beyond what has been achieved to chart the territories yet to be explored, where the fusion of symmetry discovery and unfolding might lead us next.

I.H.1 Beyond Linear Symmetries

The restriction to linear equiareal maps, while yielding impressive results, also represents a glaring blind spot in our analyses. Nature's symmetries extend far beyond the affine group, encompassing transformations that twist, fold, and reshape phase space in ways that linear maps cannot capture. Consider the Hénon map, a deceptively simple area-preserving transformation,

$$g(x, y) = (x, y - x^2) \tag{I.42}$$

It is but one of a vast and relatively uncharted space of the full (nonlinear) *equiareal group*. The structure of the equiareal group is rich enough to allow for a rich theory of symmetries and their associated invariances with the right computational tools. In particle physics, such transformations emerge naturally, and are ubiquitous, from the

rapidity transformation in relativistic kinematics, to the symplectic structure imposed by Poisson brackets on classical phase space and the correponsing commutation relations of quantum mechanics. Normalizing flows, on account of their bijective nature, could be a natural fit for exploring symplectic geometries and their associated symmetries. That said, the greater challenge lies not in the architecture but in the search space. Linear transformations form a finite dimensional Lie group, amenable to systematic exploration. Nonlinear symplectic maps form an infinite dimensional space, requiring clever parametrizations and constraints. Recent work on Hamiltonian neural networks**[cite –KD]** suggests one possible path to parametrize transformations through generating functions that automatically preserve symplectic structure.

A few examples of particle physics applications of nonlinear symmetries are

- Jet substructure: The Lund plane reveals approximate scale invariant symmetries best expressed through logarithmic transformations**[cite –KD]**
- Heavy flavor physics: Dalitz plot analyses exhibit nonlinear symmetries from resonance dynamics**[cite –KD]**
- Multiparticle correlations: Collective flow in heavy ion collisions follows nonlinear hydrodynamic symmetries**[cite –KD]**

The path forward requires synergy between differential geometry, machine learning, and physics intuition, a synthesis that will push the boundaries of all three fields.

I.H.2 Approximate Symmetries and Symmetry Breaking

Perfect symmetries, while mathematically elegant, rarely survive contact with experimental reality. The universe abounds in approximate symmetries—patterns that almost hold, invariances with small violations, conservation laws with miniscule deviations. Understanding and exploiting these near–symmetries represents perhaps the most important frontier for practical applications.

An approximate symmetry is a transformation under which probability distributions remain invariant up to small, controlled deviations, often characterized by a symmetry–breaking parameter $\epsilon \ll 1$.

The mathematical framework for approximate symmetries builds on perturbation theory. For a transformation g_ϵ depending on breaking parameter ϵ,

$$p(g_\epsilon(x))|g'_\epsilon(x)| = p(x) + \epsilon\Delta p(x) + O(\epsilon^2) \tag{I.43}$$

The first–order breaking term $\Delta p(x)$ encodes how symmetry violation depends on phase space location, crucial information for precision measurements.**[cite –KD]**

Consider isospin symmetry in nuclear physics—beautiful at low energies, progressively violated as electroweak effects become important. A symmetry–aware unfolding framework must gracefully handle this energy–dependent breaking:

$$\mathcal{L}_{\text{approx}} = \mathcal{L}_{\text{data}} + \lambda(E)\mathcal{R}_{\text{isospin}} \tag{I.44}$$

where $\lambda(E) \propto \exp(-E/\Lambda_{\text{QCD}})$ weakens the constraint at high energies.**[cite –KD]**

The soft symmetry breaking framework offers particular promise. Instead of binary classification (symmetric or not), one could attempt to quantify the degree of symmetry through continuous measures,

$$S(p, g) = 1 - \frac{D_{KL}(p||g \cdot p)}{\log 2} \tag{I.45}$$

This symmetry score $S \in [0, 1]$ would enable gradient–based optimization while naturally handling approximate invariances.**[cite –KD]**.

I.H.3 A unified framework

As we stand at the confluence of symmetry discovery and unfolding, we glimpse the outlines of something larger—a unified framework where measurement, inference, and physical understanding merge into a coherent self-consistent cycle of inference. Symmetries discovered from data improve unfolding precision, and better unfolding reveals cleaner symmetry patterns. Additionally, every symmetry provides a consistency check. Violations of expected symmetries signal systematic uncertainties. Finally, the discovered symmetries themselves are physics outputs. The questions of which symmetries hold and where they break are not only of applied interest, but also of fundamental theoretical interest to our of the laws of physics.

Looking ahead, several concrete developments seem within reach including automated analysis pipelines that start with raw detector data, automatically discover

relevant symmetries, perform symmetry-aware unfolding, quantify uncertainties, all in a self-consistent unbinned chain of inference. The marriage of machine learning's pattern discovery with physics' principled reasoning promises a new era of precision measurement, where every symmetry discovered is both a practical tool and a theoretical insight.