Mind the Gap: Navigating Inference with Optimal Transport Maps

Malte Algren (malte.algren@unige.ch), **Tobias Golling** (tobias.golling@unige.ch)
University of Geneva

Francesco Armando Di Bello (francescoarmando.dibello@unige.it)
University of Pisa

Christopher Pollard (christopher.pollard@warwick.ac.uk)
University of Warwick

Abstract

Machine learning (ML) techniques have recently enabled enormous gains in sensitivity to new phenomena across the sciences. In particle physics, much of this progress has relied on excellent simulations of a wide range of physical processes. However, due to the sophistication of modern machine learning algorithms and their reliance on high-quality training samples, discrepancies between simulation and experimental data can significantly limit their effectiveness. In this work, we present a solution to this "misspecification" problem: a model calibration approach based on optimal transport, which we apply to high-dimensional simulations for the first time. We demonstrate the performance of our approach through jet tagging, using a dataset inspired by the CMS experiment at the Large Hadron Collider. A 128-dimensional internal jet representation from a powerful general-purpose classifier is studied; after calibrating this internal "latent" representation, we find that a wide variety of quantities derived from it for downstream tasks are also properly calibrated: using this calibrated high-dimensional representation, powerful new applications of jet flavor information can be utilized in LHC analyses. This is a key step toward allowing the unbiased use of "foundation models" in particle physics. More broadly, this calibration framework has broad applications for correcting high-dimensional simulations across the sciences.

1 Introduction

In hierarchical models, the likelihood of observing some datum x, given a parameter of interest θ, is determined by marginalizing over any stochastic latent variable ω: $p(x|\theta) = \int d\omega\, p(x|\omega,\theta)\, p(\omega|\theta)$. An important subset of hierarchical models comprises those which factorize, such that x only depends on θ through ω, i.e. $p(x|\omega,\theta) = p(x|\omega)$. In such situations, it may be possible to calculate $p(x|\omega)$ separately, or to constrain it via auxiliary experiments. Indeed, this is often the motivation for constructing the likelihood $p(x|\theta)$ in terms of a latent ω: when $p(x|\theta) = \int d\omega\, p(x|\omega)\, p(\omega|\theta)$, then both $p(x|\omega)$ and $p(\omega|\theta)$ can be independently derived or verified.

One method to correct to the simulated density $p_{\text{sim}}(x|\omega)$ is to derive the conditional transport function T_ω (and corresponding transport map $(T_\omega)_\#$) Pollard and Windischhofer [2022], ATLAS Collaboration [2025a], Bunne et al. [2023] such that

$$(T_\omega)_\#\, p^{\text{CR}}_{\text{sim}}(x|\omega) \equiv p^{\text{CR}}_{\text{sim}}(T_\omega\, x|\omega) = p^{\text{CR}}_{\text{data}}(x|\omega). \tag{1}$$

Machine Learning and the Physical Sciences Workshop, NeurIPS 2025.

in some calibration region (CR). Inference on θ is then performed on observations in the "signal region" (SR) resulting in the corrected likelihood

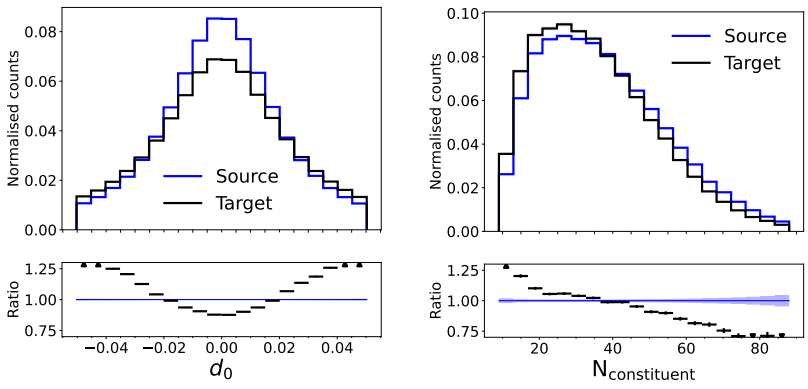
$$p(x|\theta) = \int d\omega \, (T_\omega)_\# \, p_{\text{sim}}(x|\omega) \, p_{\text{sim}}(\omega|\theta). \tag{2}$$

This transport map moves simulated observations based on what is observed in the calibration region. The suitable transport map for correcting simulations is the *optimal* one, i.e. the one which minimally alters $p_{\text{sim}}(x|\omega)$ for each choice of ω while still closing with $p_{\text{data}}(x|\omega)$. The optimal transport (OT) map is the solution to this closure problem that is the least invasive, in the sense that it preserves the full structure of $p_{\text{sim}}(x|\omega)$ as well as possible.

2 Simulation

Two datasets have been generated based on `JetClass` Qu et al. [2022], a well-established benchmark for jet classification tasks in HEP. They are generated using standard Monte Carlo event generators widely adopted in LHC experiments and incorporate detector response and reconstruction effects simulated with `DELPHES` de Favereau et al. [2014] (v3.4.2).

The first dataset, referred to as the *source*, is produced using the standard `JetClass` configuration. The second dataset, referred to as the *target*, is created by modifying the smearing functions of the `DELPHES` detector simulation and the renormalization and factorization scales of the QCD multijet events are increased from 1 to 2. Fig. 1 compares the distributions of the transverse impact parameter (d_0) and the number of jet constituents for the *source* and *target* datasets.



(a) (b)

Figure 1: The distributions are represented as follows: source distribution (blue) and target distribution (black). (a) shows the change in the transverse impact parameter (d_0), while (b) illustrates the change in number of constituents ($N_{constituent}$) between the datasets.

3 Method

Constructing the Classifier

Modern methods for tackling the jet tagging challenge consists of transformer-based classifiers Vaswani et al. [2017], ATLAS Collaboration [2025b], CMS Collaboration [2024]. We designed a transformer-based classifier that estimates the conditional probability of jet origin, $p(l \mid \{\vec{c}_{\text{obs}}\}, \text{jet}_{\text{obs}})$, where $\vec{c}_{\text{obs}}$ represents jet constituent features and jet_{obs} includes jet-level observables such as p_T, η, and E. A dedicated *ClassToken* aggregates information of the jet into a 512-dimensional latent representation.

Afterwards this latent representation is passed through a multi-layer perceptron head (MLP Head), which outputs a 10-dimensional vector of unnormalized class scores (logits). The first 128-dimensional layer of the MLP Head (z_{128}) is the target for calibration, while subsequent layers

are used for evaluation [1]. A diagram of the model architecture is shown in Fig. 2. The model is trained on the `JetClass` dataset Qu et al. [2022].

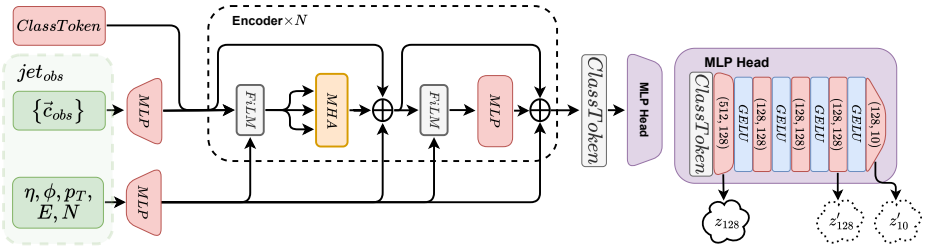
Deriving the Optimal Transport Map

The calibration aims to derive an optimal transport map, $\widehat{T}_\omega$, that aligns the source latent distribution $p_{\text{sim}}(z_{128} \mid \omega)$ with the target distribution $p_{\text{data}}(z_{128} \mid \omega)$. Here, ω represents jet kinematic features, though this work focuses solely on the q/g label.

The transport map $\widehat{T}_\omega$ is learned by minimizing the Euclidean transport cost between the source and target distributions Korotin et al. [2021], Makkuva et al. [2019] using input-convex neural networks (ICNNs) Amos et al. [2016]. The optimization problem is formalized as:

$$W_2^2(z, y) = \sup_{f(y) \in \text{cvx}(y)} \inf_{g(z) \in \text{cvx}(z)} f(\nabla_z g(z)) - \langle z, \nabla_z g(z) \rangle - f(y), \tag{3}$$

where f and g are ICNNs, and $z \sim p_{\text{sim}}$, $y \sim p_{\text{data}}$. At convergence, the optimal transport map is given by $\widehat{T}_\omega(z|\omega) = \nabla_z g(z|\omega)$, mapping source latent representations to calibrated ones.



(a) (b)

Figure 2: The two diagrams show (a) the transformer encoder and (b) the output multi-layer perceptron head (MLP Head). The transformer architecture incorporates a sequence of encoder layers Xiong et al. [2020], each comprising multi-head self-attention mechanisms (MHA) Vaswani et al. [2017] followed by feed-forward neural networks (MLP) Shazeer [2020] with Feature-wise Linear Modulation (FiLM) Perez et al. [2017], Peebles and Xie [2023] for conditional integration. The MLP Head consists of a sequence of linear transformations and GELU Hendrycks and Gimpel [2016] with a final output size of 10 for the class probabilities.

The architecture and training of the ICNNs follow Ref. Huang et al. [2021] and the procedure of Ref. Algren et al. [2024], which has been shown to be resilient to overfitting. Specifically, we use 4 ICNN blocks with depth 2 and a width of 2048 neurons per layer, using the zeroed softplus activation function ATLAS Collaboration [2025a]. Training uses the AdamW optimizer Loshchilov and Hutter [2017], we alternate optimization between f and g, performing four updates of f and ten updates of g per training step, for a total of 512,000 steps. We implemented both the training loop and ICNN models in PyTorch Paszke et al. [2019].

4 Results

To visualize the impact of calibration within the high-dimensional latent space z_{128}, Principal Component Analysis (PCA) Shlens [2014] is applied to reduce the dimensionality to the five leading principal components [2]. The distributions of these components are presented as corner plots in Fig. 3.

Fig. 3(a) illustrates the presence of the expected discrepancies between the source and target data distributions. After applying the optimal transport map, Fig. 3(b) shows a significant reduction in these discrepancies, indicating successful alignment between the latent representations of the two domains.

[1] The choice of 128 dimensional latent space is arbitrary, but we thought it would provide a good balance between expressiveness and computational efficiency.

[2] The first five principal components are chosen for visual convenience and retain 87% of the variance.

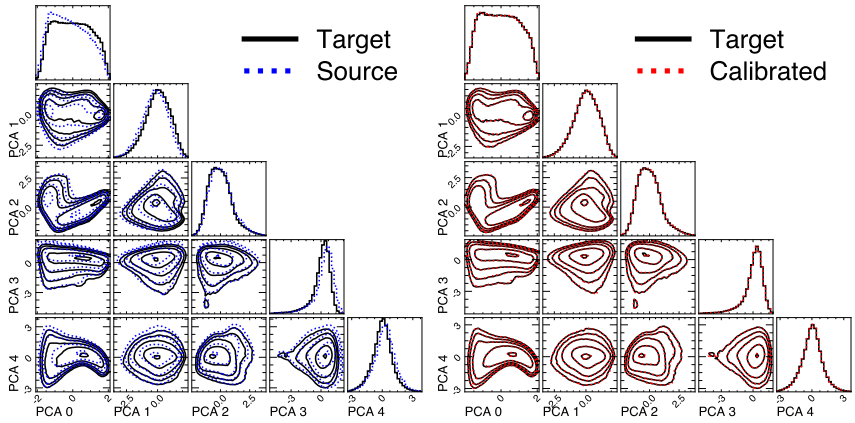
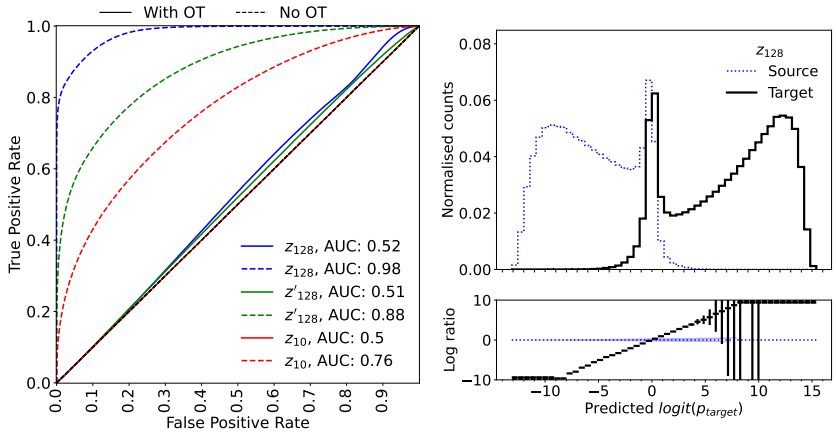


Figure 3: Corner plots of the first five principal components derived from PCA applied to the z_{128} latent space. The 2D contours show 5%, 50%, 80%, 90%, and 95% percentiles. (a) Comparison between the target distribution (black) and the source distribution (blue). (b) Comparison between the target distribution (black) and the calibrated distribution (red).



(a) (b)

Figure 4: (a) Receiver operating characteristic (ROC) curves for discriminators trained to differentiate between source and target distributions across three representational spaces: the 128-dimensional latent space (z_{128}), the final 128-dimensional layer in the MLP Head (z'_{128}), and the 10-dimensional output space (z'_{10}). The discriminators are subsequently evaluated on the calibrated versus target distributions, (b) Marginal distribution of the discriminator scores in z_{128}, between the source (blue) and target (black) distributions.

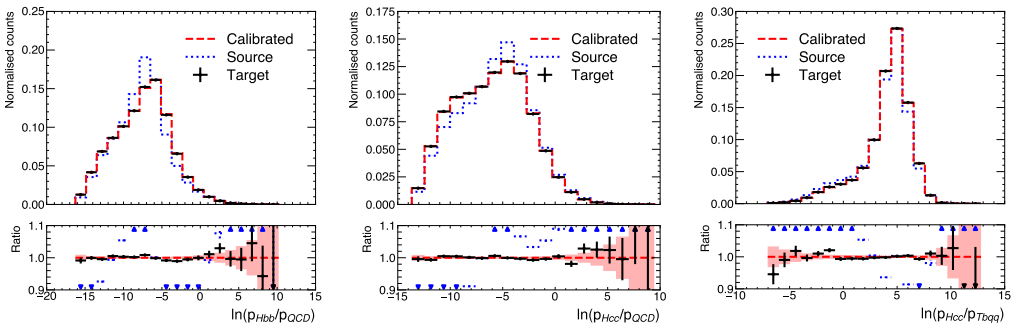
The calibration is also evaluated using discriminators trained on different layers of the MLP Head. Before calibration, source vs target discrepancies were largest in the initial 128-dimensional latent space (z_{128}) and progressively smaller in deeper layers, nearly vanishing at the output layer (z'_{10}).

After calibration, using the same source vs target discriminators[3], they could not distinguish between calibration and target in z'_{128} and z'_{10}, and with minor discrepancies in z_{128}. These metrics can be seen in Fig. 4(a).

Discriminators between calibration and target have also been trained. In z_{128}, there was a strong discrepancy (AUC ≈ 1). However, in the penultimate and output layer, the AUC≈ 0.5 meaning that the mismodeling introduced in z_{128} does not propagate to the output layer.

[3]So the discriminators are not trained between calibration vs target.

To validate the calibration's effect on the output space, the log-likelihood discriminants are analyzed in Fig. 5, with Higgs decays as signal and multijet/top events as background. Before calibration, source and target distributions showed clear discrepancies between the two, which have been removed by the calibration.



(a) (b) (c)

Figure 5: Comparisons between the marginal distributions of the physics-motivated one-dimensional discriminants. The distributions are represented as follows: target distribution (black), calibrated distribution (red), and source distribution (blue). The calibration is derived on the z_{128} latent space, and then the subsequent MLP Head layers are applied to map it to the output space. Afterwards the output space is projected onto physics-motivated one-dimensional discriminants.

5 Conclusion

A calibration strategy based on optimal transport has been presented, targeting latent representations within transformer-based classifiers used in high-energy particle physics. By applying the calibration directly in the high-dimensional latent space, rather than on low-dimensional observables or classifier outputs, the proposed method enables more flexible and detailed correction of mismodeling effects present in simulated data. This approach also establishes a foundation for the direct calibration of large, pre-trained models in physics applications.

Quantitative evaluations using discriminator networks indicate that the calibrated latent representations achieves very good agreement with the target domain for physics-relevant inference tasks. While this work focuses on jet classification, the methodology is broadly applicable to a range of reconstruction tasks in high-energy physics that utilize deep learning models, and, we believe, throughout the sciences.

The size of the latent space was arbitrarily chosen to be 128 dimensions for this study, however, future work could explore at which dimensionality can a jet be optimally represented.

Further investigations into the non-closures in the space of the calibration are warranted, as ideally the calibration should not introduce new discrepancies. It is also critical to assess the optimality of the calibrated discriminants. Both are ongoing subjects of research.

References

Malte Algren, John Andrew Raine, and Tobias Golling. Decorrelation using optimal transport. *The European Physical Journal C*, 84(6), June 2024. ISSN 1434-6052. doi: 10.1140/epjc/s10052-024-12868-6. URL `http://dx.doi.org/10.1140/epjc/s10052-024-12868-6`.

Brandon Amos, Lei Xu, and J. Zico Kolter. Input convex neural networks. *CoRR*, abs/1609.07152, 2016. URL `http://arxiv.org/abs/1609.07152`.

ATLAS Collaboration. A continuous calibration of the atlas flavour-tagging classifiers via optimal transportation maps. 2025a. URL `https://arxiv.org/abs/2505.13063`.

ATLAS Collaboration. Transforming jet flavour tagging at ATLAS. *arXiv preprint*, May 2025b. doi: 10.48550/arXiv.2505.19689.

Charlotte Bunne, Andreas Krause, and Marco Cuturi. Supervised training of conditional monge maps, 2023. URL https://arxiv.org/abs/2206.14262.

CMS Collaboration. A unified approach for jet tagging in Run 3 at $\sqrt{s} = 13.6$ TeV in CMS. *CMS-DP-2024-066*, 2024. URL https://cds.cern.ch/record/2904702.

J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi. Delphes 3: a modular framework for fast simulation of a generic collider experiment. *Journal of High Energy Physics*, 2014(2), February 2014. ISSN 1029-8479. doi: 10.1007/jhep02(2014)057. URL http://dx.doi.org/10.1007/JHEP02(2014)057.

Dan Hendrycks and Kevin Gimpel. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *CoRR*, abs/1606.08415, 2016. URL http://arxiv.org/abs/1606.08415.

Chin-Wei Huang, Ricky T. Q. Chen, Christos Tsirigotis, and Aaron Courville. Convex potential flows: Universal probability distributions with optimal transport and convex optimization, 2021. URL https://arxiv.org/abs/2012.05942.

Alexander Korotin, Lingxiao Li, Aude Genevay, Justin Solomon, Alexander Filippov, and Evgeny Burnaev. Do neural optimal transport solvers work? A continuous wasserstein-2 benchmark. *CoRR*, abs/2106.01954, 2021. URL https://arxiv.org/abs/2106.01954.

Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017. URL http://arxiv.org/abs/1711.05101.

Ashok Vardhan Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason D. Lee. Optimal transport mapping via input convex neural networks. *CoRR*, abs/1908.10962, 2019. URL http://arxiv.org/abs/1908.10962.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Z. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703, 2019. URL http://arxiv.org/abs/1912.01703.

William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. URL https://arxiv.org/abs/2212.09748.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual Reasoning with a General Conditioning Layer, December 2017.

Chris Pollard and Philipp Windischhofer. Transport away your problems: Calibrating stochastic simulations with optimal transport. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 1027:166119, March 2022. ISSN 0168-9002. doi: 10.1016/j.nima.2021.166119. URL http://dx.doi.org/10.1016/j.nima.2021.166119.

Huilin Qu, Congqiao Li, and Sitian Qian. Jetclass: A large-scale dataset for deep learning in jet physics. June 2022. doi: 10.5281/zenodo.6619768. URL https://doi.org/10.5281/zenodo.6619768.

Noam Shazeer. Glu variants improve transformer, 2020. URL https://arxiv.org/abs/2002.05202.

Jonathon Shlens. A tutorial on principal component analysis. *CoRR*, abs/1404.1100, 2014. URL http://arxiv.org/abs/1404.1100.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture. *CoRR*, abs/2002.04745, 2020. URL https://arxiv.org/abs/2002.04745.