

Exploring the synergy of kinematics and dynamics for collider physics

Kayoung Ban^{1,2,*} Kyoungchul Kong^{3,†} Myeonghun Park^{1,2,‡} and Seong Chan Park^{1,2,§}

¹*Department of Physics and IPAP, Yonsei University, Seoul 03722, Republic of Korea*

²*School of Physics, Korea Institute for Advanced Study, Seoul 02455, Korea*

³*Department of Physics and Astronomy, University of Kansas, Lawrence, Kansas 66045, USA*

⁴*School of Natural Sciences, Seoultech, Seoul 01811, Republic of Korea*



(Received 28 December 2023; accepted 20 November 2024; published 24 December 2024)

In collider experiments, an event is characterized by two distinct yet mutually complementary features: the “global features” and the “local features.” Kinematic information such as the event topology of a hard process, masses, and spins of particles comprises global features spanning the entire phase space. This global feature can be inferred from reconstructed objects. In contrast, representations of particles in gauge groups, such as quantum chromodynamics (QCD), offer localized features revealing the dynamics of an underlying theory. These local features, particularly observed in the patterns of radiation as raw data in various detector components, complement the global kinematic features. We propose a simple but effective neural network architecture that seamlessly integrates information from both kinematics and QCD to enhance the signal sensitivity at colliders.

DOI: [10.1103/PhysRevD.110.115035](https://doi.org/10.1103/PhysRevD.110.115035)

I. INTRODUCTION

The Large Hadron Collider (LHC) has ushered in an era of data proliferation, characterized by unprecedented complexity and volume, which accelerated new physics searches beyond the Standard Model. Amidst this abundance of data, deep neural networks (DNNs) have risen as formidable and essential tools [1–6], adept at unraveling the intricate correlations veiled within the vast, multidimensional data emanating from particle collisions [7–10].

In collider phenomenology, kinematic features are characterized by global stance as depicted through various observables such as invariant mass, angular distance, and transverse momentum of final state particles encapsulating the event characteristics. They capture the comprehensive dynamics of particle collisions, offering insights, and intrinsic attributes reflective of the high-energy phenomena involved [10–12]. Concurrently, local features, in particular, come from activities of QCD interactions and capture the aspects of SU(3) gauge interactions. These local attributes predominantly pertain to the QCD color charges of quarks and gluons [13–18]. One often utilizes

combinations of the global (kinematic) and the local (QCD specific) features in analyzing a collision event at the LHC, which is illustrated in Fig. 1.

To effectively combine these data types, a multimodal neural network has been utilized, which generally employs two main approaches. The first approach is focused on constancy, identifying correlations within the data to understand how different data types relate and influence each other, thereby enhancing the network’s predictive accuracy through consistent patterns across modalities [19–22]. The second approach centers on complementarity, acting on independent information from each data type to provide a more complete picture [23–26]. This method enriches the overall analysis by integrating these distinct yet complementary data points, leading to a more nuanced understanding. In the research area of high-energy physics, we adopt a complementary approach rather than consistency to maximize the use of information. The optimization of combinations through complementarity is challenging, as it often sensitively depends on the specific data set and underlying physics, an aspect not significantly emphasized in previous studies [27–29].

In this paper, we propose a novel approach integrating kinematic variables and QCD features in a multimodal neural network in a complementary way. The focus is to investigate and mitigate the unintended shadowing effects that can occur in multimodal deep neural networks when integrating local and global features. Our architecture is structured to maintain the fidelity of both kinematic variables and QCD features, ensuring that each data type is complemented and without shadowing. To navigate the intricate interplay between the different types of particle

*Contact author: kban@kias.re.kr

†Contact author: kckong@ku.edu

‡Contact author: parc.seoultech@seoultech.ac.kr

§Contact author: sc.park@yonsei.ac.kr

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article’s title, journal citation, and DOI. Funded by SCOAP³.

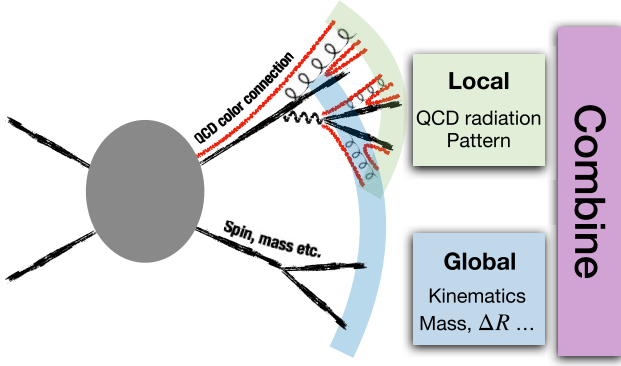


FIG. 1. Schematic overview of the types of information at a collider. A meticulous combining methodology is crucial, preserving and leveraging independent features to understand an underlying physics.

data—encompassing global kinematic variables to local QCD features—a model capable of handling such complexity with precision is necessary.

II. OPTIMIZING COMBINATION

To consider different types of information in a multimodal network, we take a one-view-one-network strategy which assigns a separate neural network for each type of information [30]. To handle with information from kinematics, we assign a fully connected network (FCN) that takes multidimensional vector-type data of kinematic variables as input. For the information from QCD activities, here energy deposits from a QCD color flow, we utilize a convolutional neural network (CNN) that takes charged particles and neutral particles from particle flow information [31] as separated channels of a pixelated image in the (η, ϕ) plane. Here we emphasize that one can choose a network model for each branch of different information depending on the type of input data. For example, one can utilize a Lorentz boost network with reconstructed particles as input vectors [32] instead of a FCN with kinematic variables, and a transformer based network with low-level particles for QCD information [33] instead of applying a CNN. But most studies with low-level particles have been focused on the boosted jet analysis where information is localized within the size of a boosted-jet and the number of particles under consideration is $\mathcal{O}(10) - \mathcal{O}(100)$. In our case, the average number of particles within the range of $|\eta| < 2.5$ in the (η, ϕ) plane is $\mathcal{O}(250)$. Thus we adopt the CNN for pixelization and convolutions to reduce computational complexity efficiently.

In short, we employ the following three models as a comparative set for our study:

- (a) FCN with kinematic variables
- (b) FCN + CNN with QCD information (1) without and (2) with an attention layer

- (c) FCN + CNN with QCD information after Riemannian preprocessing¹ (1) without and (2) with an attention layer

We provide schematic diagrams of models (1) without and (2) with the attention layer in Fig. 2. The FCN model (a) solely considers kinematic variables. It comprises four fully connected layers: each of which utilizes ReLU activation and is interspersed with batch normalization. The model is completed with an output layer with a sigmoid activation function. Models (b) and (c), as multimodal designs, will be detailed in the following.

CNN branch (local features): The CNN branch accepts 2D arrays of shape $32 \times 32 \times 2$ as input. Two channels consist of images of charged particles and neutral particles, respectively. This is in accordance with the previous study [28].

FCN branch (global features): The FCN branch takes a 1D array of N kinematic variables as inputs.

Merging: In the previous studies [28,29,34–36], information from kinematics and QCD is simply concatenated in classification models. However, simply concatenating the outputs from the FCN and CNN branches often results in an imbalance. This concatenation approach often results in a dominance of kinematic features derived from the FCN, which can inadvertently overshadow the valuable contributions of local features, such as QCD color structures. Also, multimodal networks generally do not apply activation in the last layer of each branch to prevent loss of information. However, this can cause the extracted layer to be an imbalanced data scale. To resolve these issues, we introduce additional network structures.

Attention layer: To mitigate this imbalance due to the nonactivated layer, and we employ regularization and an attention mechanism into a multimodal framework. Specifically, an L2 regularization term is added, governed by a hyperparameter l_2 , to the loss function in the output layer of each individual model branch. This constraint is essential for preventing any single branch from dominating the multimodal model, thereby ensuring that both local and global features contribute optimally to the final prediction. This penalizes large weights, thereby encouraging the multimodal model to merge as balanced with the two different types of model. The L2 regularization term has the form of a Lagrangian multiplier as

$$\text{Regularization term} = l_2 \times \sum_{k=1}^M W_k^2. \quad (1)$$

W_k denotes the weights of the output of the FCN or CNN branch, and l_2 is a hyperparameter for a regularization.

To resolve the overshadowing problem, we define the attention score as a with the softmax function [24,37–39]

¹We will introduce the Riemannian preprocessing applied in our study in the following Sec. III C.

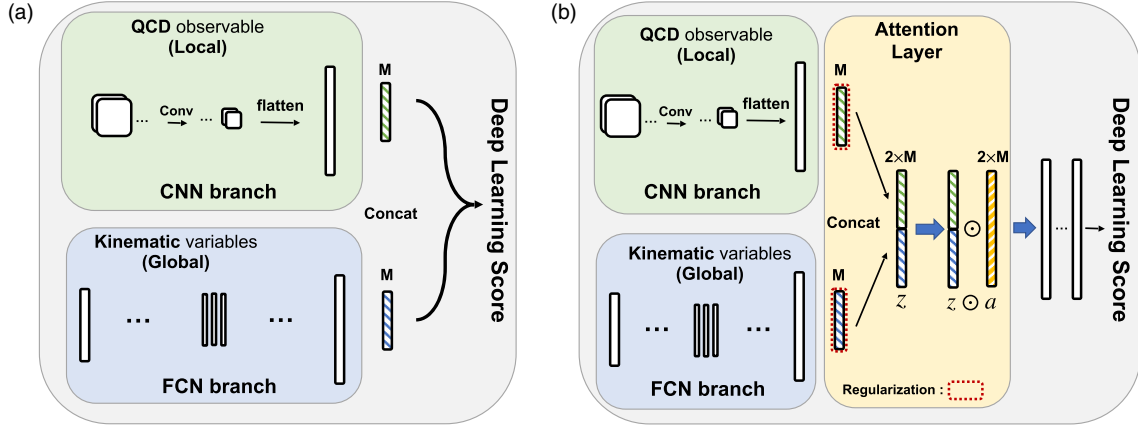


FIG. 2. Schematic plots for neural network structures: (a) conventionally used one in previous studies only with concatenation and (b) our proposed one in which we add a L2-regularized attention mechanism in a concatenation procedure.

$$a(z) = \frac{e^{f_{\text{attn}}(z)}}{\sum_{j=1}^{2M} e^{f_{\text{attn}}(z_j)}}, \quad (2)$$

where the $z = [z_{1:M}:z_{M+1:2M}]$ is the concatenated layer from the CNN and FCN, respectively, and $f_{\text{attn}}(x)$ is a trainable linear transformation, $W_{\text{attn}}x + b_{\text{attn}}$. Attention values are calculated as the element-wise product ($z \odot a$) between the attention score and the concatenated layer. With above additions, our neural network can learn how to choose important information and it becomes more transparent by showing its focus on each branch in classifying a signal and a background. The attention values are then connected with a fully connected layer for classification.

III. EXAMPLE: HH vs $t\bar{t}$

With the current interests of the LHC is being focused on the “precision Higgs,” we consider a process $pp \rightarrow HH$ at the LHC [28,40–49]. This double Higgs production channel offers a unique probe into the Higgs boson self-coupling which is a key parameter in the Higgs potential. Accurate measurements of the Higgs self-coupling provide insights on electroweak symmetry breaking, stability of the electroweak vacuum, (critical) Higgs inflation, and the potential for new physics beyond the Standard Model (SM) [50–53].

Out of various decay modes in the double Higgs channel, we consider the $(b\bar{b}\ell\bar{\ell}\nu\bar{\nu})$ mode from $H \rightarrow b\bar{b}$ and $H \rightarrow WW^* \rightarrow \ell\bar{\nu}\ell\nu$. The dominant background is $t\bar{t}$ as an irreducible one. This example serves as an exemplary case study for two reasons: firstly, the kinematic characteristics of the signal and backgrounds exhibit substantial differences; and secondly, the QCD radiation patterns from the $b\bar{b}$ -system in both the signal and backgrounds offer distinctive features attributed to the disparate $SU(3)$ representations of the H and the top quark.

A. Event selection

To generate Monte Carlo samples, we use the standard chain of MadGraph5_aMC@NLO [54], PYTHIA8 [55], and

DELPHES [56] for the 14 TeV LHC. As our focus is to optimize the combination of various information, we do not consider contaminations from pile-up. The expected number of pile-ups in the HL-LHC is approximately 200 per bunch crossing in ATLAS [57], and various algorithms utilizing machine learning (ML) have been proposed and applied to the LHC analyses [58–62] which do not interfere with an analysis process and a ML architecture used here. Jet reconstruction is performed based on the anti- k_T algorithm [63] with a radius parameter $\Delta R = 0.4$ and a transverse momentum threshold $p_{Tj} > 20$ GeV. The b -tagging efficiency is set at $\epsilon_{b \rightarrow b} = 0.7$, with a misidentification rate of $\epsilon_{c \rightarrow b} = 0.2$ and $\epsilon_{j \rightarrow b} = 0.01$. The baseline cuts are applied as follows. The transverse momentum p_T of the two leading b -tagged jets is set to be greater than 30 GeV, and two isolated leptons of opposite charge are required to have p_T greater than 20 GeV. The missing transverse momentum should exceed 20 GeV. The separation in the $\eta - \phi$ space for $b\bar{b}$ pairs, represented by $\Delta R_{b\bar{b}}$, is restricted to be less than 1.8, while for lepton pairs, $\Delta R_{\ell\bar{\ell}}$ should be under 1.3. The invariant mass of the lepton pairs $m_{\ell\bar{\ell}}$ is constrained to be less than 65 GeV and the invariant mass of b -tagged jets lies between $95 \text{ GeV} \leq m_{b\bar{b}} \leq 140 \text{ GeV}$. Lastly, the pseudorapidity η_j of the jets is capped at 2.5.

B. Kinematic information

We utilize conventional $N = 10$ kinematic variables [28,64]:

- (i) The magnitude of the missing transverse momentum,
- (ii) Transverse momentum of each lepton with p_T ordering,
- (iii) The angular distance in the (η, ϕ) plane between ℓ^- and ℓ^+ , between b and $\bar{b}$,
- (iv) Invariant mass of the $(\ell^-\ell^+)$ and $(b\bar{b})$, respectively,
- (v) Transverse momentum of $(\ell^-\ell^+)$ and $(b\bar{b})$, respectively,

- (vi) The azimuthal angles between the $(\ell^-\ell^+)$ and $(b\bar{b})$.

C. QCD information and decorrelation

There is a distinct difference in the radiation patterns emanating from jets produced by color singlet and color octet particles. QCD shower radiations from a color singlet particle tend to align more closely with the direction of the other jet due to color connection and color dipole effects, resulting in soft radiation that fills the space between the two quark jets. In contrast, radiations from color octet particles tend to spread out more broadly as they are predominantly directed toward the beam axis [13,15,18].

To enhance the complementary capabilities of multi-modal deep learning, we use a Riemannian mapping which is designed to remove the characteristics of the kinematic feature $\Delta R_{b\bar{b}}$ embedded in the jet image [65]. A Riemannian projection maps the two-dimensional space of pseudorapidity (η) and azimuthal angle (ϕ) onto a Riemann sphere R_2 determined with diameter $\overline{q_1 q_2}$ with an inverse stereographic projection as depicted in Fig. 3. The energy deposit e_i within R_2 is projected in the southern hemisphere at r_i , while the projection r_j of an energy deposit e_j outside R_2 is located in the northern hemisphere. In our case study, the soft radiation e_i of $b(\bar{b})$ stemming from the color singlet (h), which exhibits color connectivity and color dipole of $b\bar{b}$ -system, is situated within the internal space between b and $\bar{b}$, and hence, it is mapped closer to the southern hemisphere of the Riemann sphere. Conversely, the radiation e_j from $b(\bar{b})$ color connected to the incoming partons via the color triplet $t(\bar{t})$ is positioned outside the great circle of R_2 and is consequently mapped closer to the northern hemisphere.

The kinematics, here p_T of the $b\bar{b}$ system, is related to the diameter of R_2 as a larger p_T (smaller $\Delta R_{b\bar{b}}$) corresponds to the smaller diameter. As difference in a color connectivity (color singlet vs nonsinglet) is only related to the angular positions of projected energy deposits on R_2 , one can focus on the color connectivity without being distracted by the kinematics with this geometric decorrelation method. We demonstrate the results of decorrelation

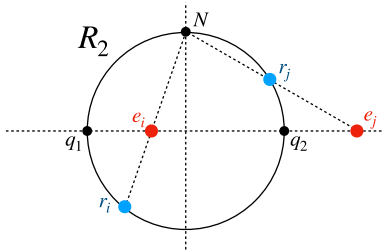


FIG. 3. Riemannian preprocessing: In the (η, ϕ) plane, the two largest energy deposits (q_1, q_2) determine a Riemann sphere R_2 with energy deposits being mapped on the R_2 . This figure is the conceptual one dimensional image of an inverse stereographic projection. More detailed information can be found in [65].

with global features using principal component analysis (PCA) and linear regression methods in the following subsection. Finally, we take a conventional procedure in dealing with jet images [14,28,34] as in Appendix A 1.

D. Decorrelating kinematic features and jet images

We describe the methodology employed to discern the degree of correlation between the kinematic features (global) and the jet images (local) of particle interactions in detail. Our procedure begins with a dimensionality reduction of the jet images, followed by linear regression to project these images into the kinematic feature space, thereby allowing us to infer the presence or absence of correlation between these two distinct feature sets. To investigate the correlation between the kinematic features (global) and the images of jet particles (local), we followed the outlined procedure. The comparative results before and after this analysis are presented in Fig. 4.

- (1) For dimension reduction of the jet image (X_{img}), which has the shape $32 \times 32 \times 2$, we applied PCA, compressing it into a 10-dimensional representation ($X_{img;PCA}$).
- (2) Using linear regression, we regressed the PCA-reduced jet image onto the kinematic feature space as

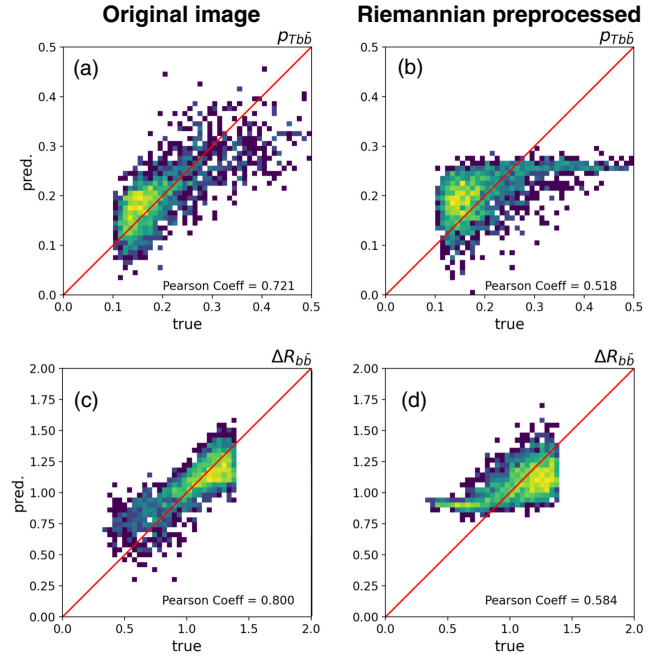


FIG. 4. Linear regression analysis between predicted and true value for (a),(b) $p_{Tb\bar{b}}$ and (c),(d) $\Delta R_{b\bar{b}}$. Each subplot corresponds to an (a), (c) original jet image and (b), (d) Riemannian preprocessed one, with the x-axis representing the true values and the y-axis depicting the predicted values from the model. The red line indicates the identity line where predicted values would exactly match the true values. Pearson correlation coefficients are displayed on each plot to provide a measure of linear correlation between the true and predicted values.

$$X_{\text{kin};\text{train}} = \mathbf{W} \times X_{\text{img};\text{train};\text{PCA}}.$$

If the regression of $X_{\text{img};\text{PCA}}$ onto X_{kin} is relatively unsuccessful, it can be indirectly inferred that there is a lack of correlation between the two feature sets.

- (3) After getting $\mathbf{W}^*$, which is optimized, results after linear regression, by using test data, we can obtain

$$X_{\text{kin};\text{predicted}} = \mathbf{W}^* \times X_{\text{img};\text{test};\text{PCA}}.$$

- (4) We then compared the distribution of the predicted kinematic features, denoted as $X_{\text{kin};\text{predicted}}$, with the actual X_{kin} .

IV. RESULTS

With an attention layer, we can assess the “importance” of each contribution in detail. We utilize the attention scores $a(z)$ of Eq. (2) from the trained multimodal model and the attention layer across two distinct scenarios: jet images and kinematic features, utilizing a test dataset of 20 samples in Fig. 5. The left plots illustrate the score distribution when L2 regularization is not applied, while the right plots display the attention scores under the influence of L2 regularization, considering both pure jet images and Riemannian preprocessed images.

The attention scores are divided with the red vertical lines along the x -axis into those originating from jet images (left of the red line) and those from kinematic features (right of the red line), each comprising $M = 16$ weights. The intensity of the colors within the heatmap corresponds to the magnitude of attention scores, indicating the focus of the model on specific features. As previously discussed, the multimodal model without L2 regularization shows a propensity for the attention score to favor kinematic features in the

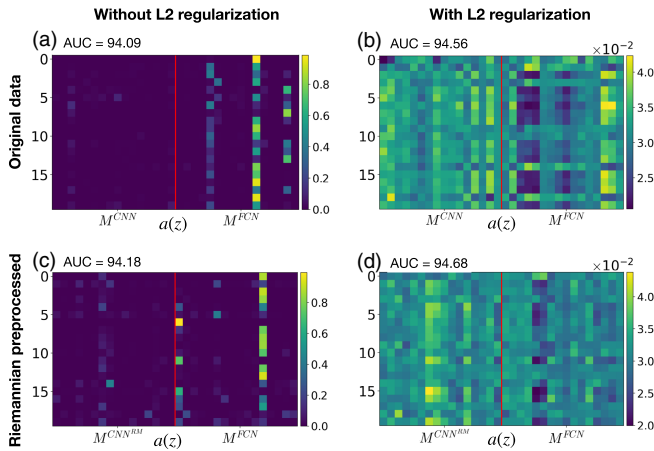


FIG. 5. Comparison of attention scores $a(z)$ in the x -axis for a test dataset comprising 20 samples with a data index in the y -axis evaluated post-training under two distinct neural network models (without L2 regularization in the left and with L2 in the right) using two different datasets (original data in the top and Riemannian preprocessed data in the bottom).

classification process. This bias likely arises from the data scale imbalance encountered during the extraction of local and global features by each branch. L2 regularization is employed to mitigate this, as shown in Fig. 5, resulting in the attention score more equitably distributing its focus between jet images and kinematic variables.

Finally, in an effort to ascertain the robustness of our models, we conducted evaluations across datasets of varying sizes, specifically those containing 5,000 (5k) and 50,000 (50k) samples. We consider the AUC (area under the curve) of the radius of curvature to gauge model performance in evaluating the power of a classification. The results of these evaluations are presented in Fig. 6, which illustrates the distribution of AUC scores for a suite of models: FCN, FCN combined with CNN without and with Riemannian mapping preprocessing, and their respective implications with an attention layer. The models equipped with attention mechanisms demonstrate an appreciable enhancement. Moreover, the consistency in AUC score distribution across the two dataset sizes for these models highlights their stability and generalizability, a desirable attribute for applications where data availability can fluctuate significantly. It is important to note that the underlying structures of the models employed in our analysis remain consistent throughout. Additionally, it is evident that when dealing with smaller training datasets such as 5k, the introduction of suitable preprocessing techniques (Riemannian mapping in our case) is crucial to ensuring the stability of the learning models. This is showcased by the improved AUC scores, indicating that the models are

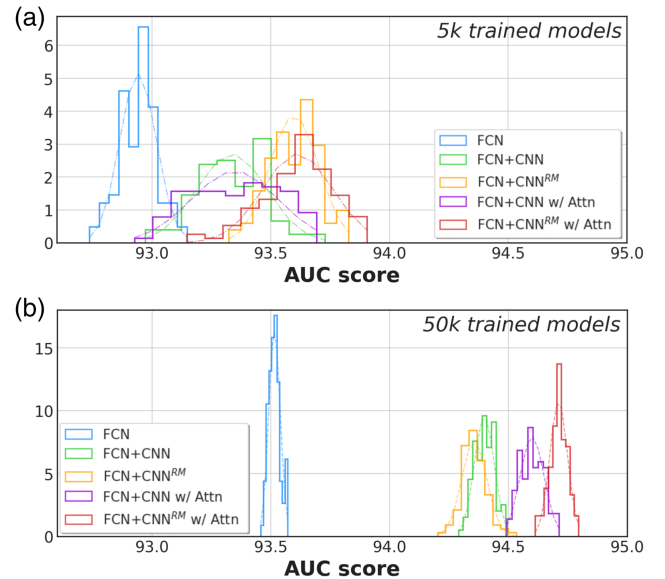


FIG. 6. Distribution of AUC scores for 100 independent trained models on datasets of (a) 5k and (b) 50k sizes. The histograms compare the AUC score performance across FCN (blue), FCN + CNN without (green) and with (yellow) Riemannian mapping preprocessing, and FCN + CNN with an attention layer without (purple) and with (red) Riemannian mapping preprocessing for both the training and testing phases.

TABLE I. Summary of efficiency (ϵ) of neural networks trained with 50,000 datasets, on a test dataset with size 50,000 samples for signal and background each.

Model	Efficiency (%)	
	ϵ_{sig}	ϵ_{bkg}
FCN 10 vars	41.28	0.160
FCN + CNN	41.27	0.144
FCN + CNN ^{RM}	41.23	0.152
FCN + CNN w/Attn	41.26	0.132
FCN + CNN ^{RM} w/Attn	41.26	0.084

more reliable and provide more consistent predictions. The hyperparameters, which are adapted in each model, are comprehensively detailed in Appendix A 2. This uniformity in model structure ensures that any observed variations in performance or outcomes can be attributed primarily to the adjustments in the hyperparameters, rather than differences in the architectural frameworks of the models.

To assess the potential of uncovering new phenomena through different neural network architectures, the discovery potential of a signal at the LHC is quantified by determining the signal significance using the likelihood ratio [66],

$$\sigma = \sqrt{-2 \ln \left(\frac{L(n_B | n_S + n_B)}{L(n_S + n_B | n_S + n_B)} \right)}, \quad (3)$$

where $L(x|n) = \frac{x^n e^{-x}}{n!}$ and n_s and n_b are the expected number of signals and background. We assume double Higgs production at the HL-LHC with an integrated luminosity of 3 ab^{-1} and select the optimal model in Fig. 6(b) to calculate the deep learning score cut efficiencies using 50,000 test samples in Table I.

We examine the changes in significance with 500 pseudoexperiments using cut efficiencies in Table I to check the significance statistics across the different pseudo experiments. In this case, we use 50 signal and 50,000 background samples and reweight the signal and background to

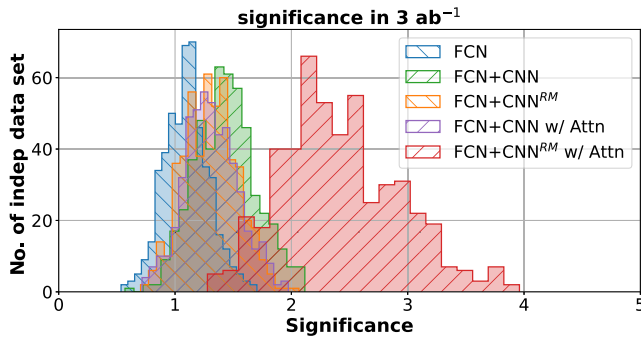


FIG. 7. Histograms showing distributions of significance σ across 500 pseudoexperiments for various neural network architectures. The significance is measured at 3 ab^{-1} with cut efficiency denoted in Table I.

TABLE II. Summary for significance σ in Fig. 7.

Model	Mean	Median	Std
FCN 10 vars	1.101	1.103	0.192
FCN + CNN	1.433	1.428	0.253
FCN + CNN ^{RM}	1.304	1.297	0.229
FCN + CNN w/Attn	1.308	1.294	0.231
FCN + CNN ^{RM} w/Attn	2.415	2.360	0.507

match the cross section. In Fig. 7 and corresponding Table II, FCN + CNN^{RM} with an attention layer also achieves consistently higher significance values compared to other models.

The results highlight that, with an ample amount of training data, it is not only the preprocessing that contributes to model performance but also the effective integration of local and global features. The combination of these features allows the model to harness comprehensive information from both the detailed local interactions and the broader global event characteristics, enhancing the ability of the model to generalize and perform robustly across varied datasets.

A. Comparison with other multimodal models

Additionally, various multimodal network structures have been proposed to address the challenges of imbalanced learning and overfitting in multimodal frameworks [67–71]. We compare the On-the-fly Gradient Modulation with the Generalization Enhancement (OGM-GE) method [68] with our attention-based multimodal model. The OGM-GE method dynamically adjusts the gradient contributions of each modality to address optimization imbalances, leading to improved performance and generalization. The OGM-GE modulates the optimization process by calculating the discrepancy ratio ρ_i^u for each modality and adjusting the gradient with a modulation coefficient k_i^u . This approach reduces the learning imbalance between the two modalities. The experimental results demonstrate that the OGM-GE method yields significant accuracy gains across various tasks, thereby confirming its utility in enhancing multimodal learning frameworks.

However, our attention-based multimodal model, integrating kinematics and color information data, demonstrated a modest performance improvement than utilizing OGM-GE as

TABLE III. Comparison of area under the ROC curve and accuracy (ACC) percentages between two multimodal models: one using FCN^{RM} + CNN with OGM-GE, and the other with attention. The models are evaluated on a dataset of 50,000 samples. For each model, the statistics are computed over 100 independent training runs.

	AUC (%)	ACC (%)
FCN + CNN ^{RM} w / OGM-GE	94.20 ± 0.05	87.28 ± 0.12
FCN + CNN ^{RM} w/Attn	94.71 ± 0.04	87.59 ± 0.09

shown in Table III. Here, we examined the statistical fluctuations of the models through 100 independent training runs, and the variance is included in the table.

V. CONCLUSION

In this work, we introduce a novel multimodal deep neural network architecture capable of synergistically incorporating both local (QCD) and global (kinematics) features relevant to collider analyses. We achieve this by adding a simple attention module in the combination stage. Here the L2 regularization plays a crucial role to increase a performance by balancing contributions from each information source. By integrating kinematic features as a source of global contextual information, the model gains a comprehensive understanding crucial for classification tasks. Simultaneously, we leverage the QCD color structure to provide local, complemented attributes that serve as complementary to the global features.

We also apply a decorrelation method using Riemannian mapping. This procedure helps extract QCD information more efficiently across a wide range of phase space, which becomes particularly important when training data is limited due to the computational cost of simulating backgrounds in the signal region.

Our results demonstrate that optimizing the combination of the two distinct features results in improved model prediction performance. This improvement shows an effective classification capability even when confronted with the inherent complexities of datasets typically encountered in particle physics experiments. Our model efficiently classifies signal/background events and adapts to different data scales and distributions, affirming its stability and versatility. Including attention mechanisms further refines the performance, particularly by emphasizing salient features and facilitating the identification of subtle yet essential correlations within the data. This suggests a promising direction for future research in applying deep learning techniques in collider physics.

We also show that our “simple” architecture is comparable to (or slightly better than) multimodal networks developed within the computer science community. As such, we expect that our architecture can be useful in other fields where multimodal networks are required. Additionally, we note that simpler neural network models are especially valuable in areas where training data is expensive to obtain.

Finally, we emphasize the universality of our method. Our multimodal network can be integrated with various neural networks for different types of information, such as the Lorentz Boost Network [32] for kinematics and ParticleNet [33] for QCD information. To optimize the combination, one simply needs to add our attention layer at the stage of merging results from the different networks.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) RS-2023-00283129 and RS-2024-00340153 (S.C.P) and NRF-2021R1A2C4002551 (M.P.). K. B. is supported in part by the KIAS Individual Grant. No. PG097601. K. K. is supported by the US DOE under Award No. DE-SC0024673.

APPENDIX: DETAILS IN A NEURAL NETWORK MODEL

1. Preparing jet images

To prepare images for CNN, we take the conventional process which centralizes $b\bar{b}$ -system. Figure 8 shows the preprocessed images.

- (1) *Image cleansing*: remove all leptons from the image.
- (2) *Particle classification*: divide the channels into two: charged particles including charged hadrons and neutral particles including photons and neutral hadrons.
- (3) *Centering*: shift the center of the image from $(0, 0)$ to $(\frac{\eta_b + \eta_{\bar{b}}}{2}, \frac{\phi_b + \phi_{\bar{b}}}{2})$ which is to the center of the reconstructed b quark pair.
- (4) *Selection*: select regions based on the b jet where $\Delta R_{b\bar{b}} > 0.1$ due to concerns that intense radiation might overshadow crucial information from soft radiation through normalization.
- (5) *Pixelization*: discretize the QCD observables which are defined in the (η, ϕ) plane to 50×50 pixels for images. We downsample the resolution of the Riemannian mapping image from 50×50 to 32×32 pixels to align with the training data size, anticipating subsequent cropping for the pure jet image.

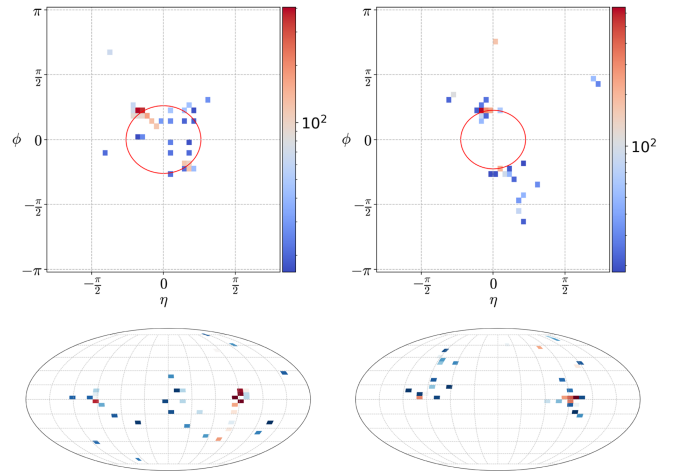


FIG. 8. Transverse momentum distribution (upper) with accumulated charged and neutral particles for a single event with hh (left) and $t\bar{t}$ (right). The red circles indicate the $\Delta R_{b\bar{b}}$ of the event. The transverse momentum distribution after Riemannian preprocessing for the same event for mollweide plane (lower).

- (6) *Normalization*: scale the pixel intensity (I_{ij}) by $I_{ij} \rightarrow I_{ij}/\Sigma I_{ij}$, where i and j are the pixel index, so that the sum of p_T in a single image is 1.
- (7) *Cropping*: For the jet image, we crop the image with (32, 32) pixels which is related to $-1.6 < \eta < 1.6$ and $-2.01 < \phi < 2.01$ for the original jet image.

The CNN branch consists of three convolutional layers, each followed by batch normalization, max-pooling, and a dropout layer with a rate of 0.3 to mitigate overfitting. Specifically, each convolutional layer employs 32 filters and a 5×5 kernel, with ReLU activation. Max-pooling is performed with a 2×2 window and a stride of 2. The output ($z_{1:M}$) is flattened and passed through a dense layer with M units without activation function.

2. Details in training a neural network model

The classification part after an attention block consists of three dense layers consists with 64 units, as indicated by the parameter M set to 16. These layers employ the ReLU activation function, and between these dense layers, batch normalization is applied to stabilize and accelerate the learning process by normalizing the activations. Finally, the network culminates in an output layer with a single unit employing a sigmoid activation function, which is standard for binary classification tasks.

For the DNN library, we use the Keras package [72] and the models are constructed with ReLU activations. We use the Adam optimizer [73] to minimize the binary cross-entropy loss function. To ensure stable training, we employ a learning rate scheduler that maintains a constant learning rate for the initial 10 epochs. Subsequently, the learning

TABLE IV. Selected hyperparameters for each model, as referenced in the results presented in Fig. 6. The table delineates the optimized values for the learning rate (ℓ) and L2 regularization coefficient (l_2) corresponding to the training datasets comprising 5,000 (5k) and 50,000 (50k) samples.

No. of training data	5k		50k	
	ℓ	l_2	ℓ	l_2
FCN 10 vars	5×10^{-3}		5×10^{-3}	
FCN + CNN	5×10^{-3}		10^{-2}	
FCN + CNN ^{RM}	$\times 10^{-2}$		10^{-2}	
FCN + CNN w/Attn	5×10^{-3}	10^{-2}	10^{-2}	10^{-2}
FCN + CNN ^{RM} w/Attn	10^{-2}	10^{-2}	5×10^{-3}	10^{-2}

rate is adjusted according to the formula $\ell \times e^{-\beta}$, where β represents the decay factor, and we set $\beta = 0.5$.

We investigate the robustness of our model with respect to varying training dataset sizes to highlight its stability. Specifically, we conduct evaluations across datasets of 5,000 and 50,000 samples. In this process, we perform a thorough optimization of parameters for each dataset and corresponding model configuration. The optimal set of hyperparameters is determined based on achieving the highest average AUC scores over 10 independent runs. We select the adjusted hyperparameters from a predefined range, with learning rates ℓ chosen from {0.01, 0.005, 0.001} and L2 regularization strengths l_2 from {0.5, 0.1, 0.05, 0.01}. The hyperparameters that result from this optimization are documented in Table IV.

-
- [1] Kim Albertsson *et al.*, Machine learning in high energy physics community white paper, *J. Phys. Conf. Ser.* **1085**, 022008 (2018).
- [2] Matthew D. Schwartz, Modern machine learning and particle physics, *Harvard Data Sci. Rev.* **3** (2021).
- [3] Matthew Feickert and Benjamin Nachman, A living review of machine learning for particle physics, *arXiv:2102.02770*.
- [4] Alexander Radovic, Mike Williams, David Rousseau, Michael Kagan, Daniele Bonacorsi, Alexander Himmel, Adam Aurisano, Kazuhiro Terao, and Taritree Wongjirad, Machine learning at the energy and intensity frontiers of particle physics, *Nature (London)* **560**, 41 (2018).
- [5] Phiala Shanahan *et al.*, Snowmass 2021 Computational Frontier CompF03 Topical Group Report: Machine learning, *arXiv:2209.07559*.
- [6] Georgia Karagiorgi, Gregor Kasieczka, Scott Kravitz, Benjamin Nachman, and David Shih, Machine learning in the search for new fundamental physics, *arXiv:2112.03769*.
- [7] Doojin Kim, Kyoungchul Kong, Konstantin T. Matchev, Myeonghun Park, and Prasanthy Shyamsundar, Deep-learned event variables for collider phenomenology, *Phys. Rev. D* **107**, L031904 (2023).
- [8] Kayoung Ban, Dong Woo Kang, Tae Geun Kim, Seong Chan Park, and Yeji Park, DeeLeMa: Missing information search with deep learning for mass estimation, *Phys. Rev. Res.* **5**, 043186 (2023).
- [9] Zhongtian Dong, Kyoungchul Kong, Konstantin T. Matchev, and Katia Matcheva, Is the machine smarter than the theorist: Deriving formulas for particle kinematics with symbolic regression, *Phys. Rev. D* **107**, 055018 (2023).
- [10] Roberto Franceschini, Doojin Kim, Kyoungchul Kong, Konstantin T. Matchev, Myeonghun Park, and Prasanthy Shyamsundar, Kinematic variables and feature engineering for particle phenomenology, *Rev. Mod. Phys.* **95**, 045004 (2023).

- [11] Tao Han, Collider phenomenology: Basic knowledge and techniques, in *Physics in $D \geq 4$* 2004 (World Scientific, Singapore, 2005), pp. 407–454, [10.1142/9789812773579_0008](#).
- [12] A. J. Barr, T. J. Khoo, P. Konar, K. Kong, C. G. Lester, K. T. Matchev, and M. Park, Guide to transverse projections and mass-constraining variables, *Phys. Rev. D* **84**, 095031 (2011).
- [13] Jason Gallicchio and Matthew D. Schwartz, Seeing in color: Jet superstructure, *Phys. Rev. Lett.* **105**, 022001 (2010).
- [14] Luke de Oliveira, Michael Kagan, Lester Mackey, Benjamin Nachman, and Ariel Schwartzman, Jet-images—deep learning edition, *J. High Energy Phys.* **07** (2016) 069.
- [15] Jason Gallicchio, John Huth, Michael Kagan, Matthew D. Schwartz, Kevin Black, and Brock Tweedie, Multivariate discrimination and the Higgs + W/Z search, *J. High Energy Phys.* **04** (2011) 069.
- [16] Michael Kagan, Image-based jet analysis, [arXiv:2012.09719](#).
- [17] Patrick T. Komiske, Eric M. Metodiev, and Matthew D. Schwartz, Deep learning in color: Towards automated quark/gluon jet discrimination, *J. High Energy Phys.* **01** (2017) 110.
- [18] Tao Han, Ian M. Lewis, Hongkai Liu, Zhen Liu, and Xing Wang, A guide to diagnosing colored resonances at hadron colliders, *J. High Energy Phys.* **08** (2023) 173.
- [19] Litian Zhang, Xiaoming Zhang, and Junshu Pan, Hierarchical cross-modality semantic correlation learning model for multimodal summarization, *Proc. AAAI Conf. Artif. Intell.* **36**, 11676 (2022).
- [20] Li Yuan, Yi Cai, Jin Wang, and Qing Li, Joint multimodal entity-relation extraction based on edge-enhanced graph alignment network and word-pair relation tagging, *Proc. AAAI Conf. Artif. Intell.* **37**, 11051 (2023).
- [21] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Jiahui Zhang, Shijian Lu, Lingjie Liu, Adam Kortylewski, Christian Theobalt, and Eric Xing, Multimodal image synthesis and editing: The generative AI era, *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 15098 (2023).
- [22] Nanyi Fei, Zhiwu Lu, Yizhao Gao, Guoxing Yang, Yuqi Huo, Jingyuan Wen, Haoyu Lu, Ruihua Song, Xin Gao, Tao Xiang *et al.*, Towards artificial general intelligence via a multimodal foundation model, *Nat. Commun.* **13**, 3094 (2022).
- [23] Jin-Young Kim and Sung-Bae Cho, A deep neural network ensemble of multimodal signals for classifying excavator operations, *Neurocomputing* **470**, 290 (2022).
- [24] Shuai Liu, Peng Gao, Yating Li, Weina Fu, and Weiping Ding, Multi-modal fusion network with complementarity and importance for emotion recognition, *Inf. Sci.* **619**, 679 (2023).
- [25] Jin-Young Kim and Sung-Bae Cho, Obfuscated malware detection using deep generative model based on global/local features, *Comput. Secur.* **112**, 102501 (2022).
- [26] Wei Tang, Fazhi He, Yu Liu, and Yansong Duan, MATR: Multimodal medical image fusion via multiscale adaptive transformer, *IEEE Trans. Image Process.* **31**, 5134 (2022).
- [27] Joshua Lin, Marat Freytsis, Ian Mould, and Benjamin Nachman, Boosting $H \rightarrow b\bar{b}$ with machine learning, *J. High Energy Phys.* **10** (2018) 101.
- [28] Jeong Han Kim, Minho Kim, Kyoungchul Kong, Konstantin T. Matchev, and Myeonghun Park, Portraying double Higgs at the large hadron collider, *J. High Energy Phys.* **09** (2019) 047.
- [29] W. Esmail, A. Hammad, and S. Moretti, Sharpening the $A \rightarrow Z^{(*)}h$ signature of the Type-II 2HDM at the LHC through advanced machine learning, *J. High Energy Phys.* **11** (2023) 020.
- [30] Xiaoqiang Yan, Shizhe Hu, Yiqiao Mao, Yangdong Ye, and Hui Yu, Deep multi-view learning methods: A review, *Neurocomputing* **448**, 106 (2021).
- [31] A. M. Sirunyan *et al.* (CMS Collaboration), Particle-flow reconstruction and global event description with the CMS detector, *J. Instrum.* **12**, P10003 (2017).
- [32] M. Erdmann, E. Geiser, Y. Rath, and M. Rieger, Lorentz boost networks: Autonomous physics-inspired feature engineering, *J. Instrum.* **14**, P06006 (2019).
- [33] Huilin Qu, Congqiao Li, and Sitian Qian, Particle transformer for jet tagging, [arXiv:2202.03772](#).
- [34] Li Huang, Su-beom Kang, Jeong Han Kim, Kyoungchul Kong, and Jun Seung Pi, Portraying double Higgs at the Large Hadron Collider II, *J. High Energy Phys.* **08** (2022) 114.
- [35] Thomas Flacke, Jeong Han Kim, Manuel Kunkel, Pyungwon Ko, Jun Seung Pi, Werner Porod, and Leonard Schwarze, Uncovering doubly charged scalars with dominant three-body decays using machine learning, *J. High Energy Phys.* **11** (2023) 009.
- [36] A. Hammad, S. Khalil, and S. Moretti, Search for mono-Higgs signals in $b\bar{b}^-$ final states using deep neural networks, *Phys. Rev. D* **107**, 075027 (2023).
- [37] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning, Effective approaches to attention-based neural machine translation, [arXiv:1508.04025](#).
- [38] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, Neural machine translation by jointly learning to align and translate, [arXiv:1409.0473](#).
- [39] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R. Hershey, Tim K. Marks, and Kazuhiko Sumi, Attention-based multimodal fusion for video description, in *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), pp. 4203–4212.
- [40] Andreas Papaefstathiou, Li Lin Yang, and Jose Zurita, Higgs boson pair production at the LHC in the $b\bar{b}W^+W^-$ channel, *Phys. Rev. D* **87**, 011301 (2013).
- [41] Jeong Han Kim, Kyoungchul Kong, Konstantin T. Matchev, and Myeonghun Park, Probing the triple Higgs self-interaction at the large hadron collider, *Phys. Rev. Lett.* **122**, 091801 (2019).
- [42] T. Huang, J. M. No, L. Pernié, M. Ramsey-Musolf, A. Safonov, M. Spannowsky, and P. Winslow, Resonant di-Higgs boson production in the $b\bar{b}WW$ channel: Probing the electroweak phase transition at the LHC, *Phys. Rev. D* **96**, 035007 (2017).
- [43] Matthew J. Dolan, Christoph Englert, and Michael Spannowsky, Higgs self-coupling measurements at the LHC, *J. High Energy Phys.* **10** (2012) 112.
- [44] Koji Nakamura, Kenji Nishiwaki, Kin-ya Oda, Seong Chan Park, and Yasuhiro Yamamoto, Di-Higgs enhancement by neutral scalar as probe of new colored sector, *Eur. Phys. J. C* **77**, 273 (2017).
- [45] A. M. Sirunyan, Search for resonant and non-resonant Higgs boson pair production in the $b\bar{b}l\bar{l}l\nu$ final state at $\sqrt{s} = 13$ TeV, *J. High Energy Phys.* **01** (2018) 054.

- [46] Albert M. Sirunyan *et al.* (CMS Collaboration), Search for resonant and nonresonant Higgs boson pair production in the $b\bar{b}\ell\ell\ell\ell$ final state in proton-proton collisions at $\sqrt{s} = 13$ TeV, *J. High Energy Phys.* **01** (2018) 054.
- [47] Morad Aaboud *et al.* (ATLAS Collaboration), Search for Higgs boson pair production in the $b\bar{b}WW^*$ decay mode at $\sqrt{s} = 13$ TeV with the ATLAS detector, *J. High Energy Phys.* **04** (2019) 092.
- [48] Georges Aad *et al.* (ATLAS Collaboration), Search for non-resonant Higgs boson pair production in the $b\bar{b}\ell\ell\ell\ell$ final state with the ATLAS detector in pp collisions at $\sqrt{s} = 13$ TeV, *Phys. Lett. B* **801**, 135145 (2020).
- [49] Georges Aad *et al.* (ATLAS Collaboration), Search for resonant and non-resonant Higgs boson pair production in the $b\bar{b}\tau^+\tau^-$ decay channel using 13 TeV pp collision data from the ATLAS detector, *J. High Energy Phys.* **07** (2023) 040.
- [50] Giuseppe Degrandi, Stefano Di Vita, Joan Elias-Miro, Jose R. Espinosa, Gian F. Giudice, Gino Isidori, and Alessandro Strumia, Higgs mass and vacuum stability in the standard model at NNLO, *J. High Energy Phys.* **08** (2012) 098.
- [51] Andrea De Simone, Mark P. Hertzberg, and Frank Wilczek, Running inflation in the standard model, *Phys. Lett. B* **678**, 1 (2009).
- [52] Yuta Hamada, Hikaru Kawai, Kin-ya Oda, and Seong Chan Park, Higgs inflation from standard model criticality, *Phys. Rev. D* **91**, 053008 (2015).
- [53] Yuta Hamada, Hikaru Kawai, Kin-ya Oda, and Seong Chan Park, Higgs inflation is still alive after the results from BICEP2, *Phys. Rev. Lett.* **112**, 241301 (2014).
- [54] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations, *J. High Energy Phys.* **07** (2014) 079.
- [55] Torbjörn Sjöstrand, Stefan Ask, Jesper R. Christiansen, Richard Corke, Nishita Desai, Philip Ilten, Stephen Mrenna, Stefan Prestel, Christine O. Rasmussen, and Peter Z. Skands, An introduction to PYTHIA8.2, *Comput. Phys. Commun.* **191**, 159 (2015).
- [56] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi (DELPHES 3 Collaboration), DELPHES 3, A modular framework for fast simulation of a generic collider experiment, *J. High Energy Phys.* **02** (2014) 057.
- [57] ATLAS Collaboration, Expected performance of the ATLAS detector under different High-Luminosity LHC conditions, CERN Report No. ATL-PHYS-PUB-2021-023, 2021.
- [58] Patrick T. Komiske, Eric M. Metodiev, Benjamin Nachman, and Matthew D. Schwartz, Pileup mitigation with machine learning (PUMML), *J. High Energy Phys.* **12** (2017) 051.
- [59] ATLAS Collaboration, Convolutional Neural Networks with Event Images for Pileup Mitigation with the ATLAS Detector, CERN Report No. ATL-PHYS-PUB-2019-028, 2019.
- [60] J. Arjona Martínez, Olmo Cerri, Maurizio Pierini, Maria Spiropulu, and Jean-Roch Vlimant, Pileup mitigation at the large hadron collider with graph neural networks, *Eur. Phys. J. Plus* **134**, 333 (2019).
- [61] Stefano Carrazza and Frédéric A. Dreyer, Jet grooming through reinforcement learning, *Phys. Rev. D* **100**, 014014 (2019).
- [62] Benedikt Maier, Siddharth M. Narayanan, Gianfranco de Castro, Maxim Goncharov, Christoph Paus, and Matthias Schott, Pile-up mitigation using attention, *Mach. Learn. Sci. Tech.* **3**, 025012 (2022).
- [63] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez, The anti- k_t jet clustering algorithm, *J. High Energy Phys.* **04** (2008) 063.
- [64] CMS Collaboration, Projected performance of Higgs analyses at the HL-LHC for ECFA 2016, CERN Report No. CMS-PAS-FTR-16-002, 2017.
- [65] Ahmed Hammad and Myeonghun Park, Riemannian data preprocessing in machine learning to focus on QCD color structure, *J. Korean Phys. Soc.* **83**, 235 (2023).
- [66] Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells, Asymptotic formulae for likelihood-based tests of new physics, *Eur. Phys. J. C* **71**, 1554 (2011); *Eur. Phys. J. C* **73**, 2501(E) (2013).
- [67] Weiyao Wang, Du Tran, and Matt Feiszli, What makes training multi-modal classification networks hard? in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 12695–12705.
- [68] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu, Balanced multimodal learning via on-the-fly gradient modulation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022).
- [69] Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li, Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition, in *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23 (Association for Computing Machinery, New York, USA, 2023), pp. 5923–5934.
- [70] Liu Yang, Zhenjie Wu, Junkun Hong, and Jun Long, MCL: A contrastive learning method for multimodal data fusion in violence detection, *IEEE Signal Process. Lett.* **30**, 408 (2023).
- [71] Yu He, Licai Sun, Zheng Lian, Bin Liu, Jianhua Tao, Meng Wang, and Yuan Cheng, Multimodal temporal attention in sentiment analysis, in *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, MuSe' 22 (Association for Computing Machinery, New York, USA, 2022), pp. 61–66.
- [72] François Chollet, Keras, <https://github.com/fchollet/keras> (2015).
- [73] Diederik Kingma and Jimmy Ba, Adam: A method for stochastic optimization, *International Conference on Learning Representations* (2014).