Amplitude Surrogates for Multi-Jet Processes

Luca Beccatini[1,2,3], Fabio Maltoni[1,2,3,4], Olivier Mattelaer[1], Ramon Winterhalder[5]

1 CP3, Université catholique de Louvain, Louvain-la-Neuve, Belgium
2 Dipartimento di Fisica e Astronomia, Università di Bologna, Italy
3 INFN, Sezione di Bologna, Bologna, Italy
4 European Organisation for Nuclear Research (CERN), Geneva, Switzerland
5 TIFLab, Università degli Studi di Milano & INFN Sezione di Milano, Milano, Italy

January 13, 2026

Abstract

Accurate and efficient amplitude predictions are essential for precision studies of multi-jet processes at the LHC. We introduce a novel neural network architecture that predicts multi-jet amplitudes by leveraging the Catani–Seymour factorization scheme and related lower-jet amplitudes, requiring the network to learn only a correction factor. This hybrid approach combines theoretical factorization with a data-driven ansatz, enabling fast and scalable amplitude predictions. Our networks also estimate the accuracy of each prediction, allowing us to selectively use results that meet a predefined accuracy threshold. In the context of leading-order event generation, this approach achieves speed-up factors of up to 20 while maintaining all observables at the percent-level accuracy.

Contents

---

1 Introduction

Accurate and efficient predictions of scattering amplitudes are essential for precision studies in high-energy physics, in particular for multi-jet processes at hadron colliders [1]. At the LHC and its upcoming high-luminosity phase, increasingly differential measurements and complex final states demand Monte-Carlo simulations that combine large event samples with high perturbative accuracy. Multi-purpose event generators such as PYTHIA8 [2], SHERPA [3], HERWIG [4], and MG5AMC [5] provide the backbone of this theoretical infrastructure by interfacing fixed-order matrix elements with parton showers, hadronization, and detector simulation.

Despite many algorithmic advances, evaluating scattering amplitudes remains a significant computational bottleneck in event generation. The cost of computing tree-level amplitudes grows rapidly with final-state multiplicity, reflecting both the combinatorial growth of Feynman diagrams and the increasing complexity of phase-space integration. While more advanced techniques [6] substantially mitigate the naïve factorial scaling, high-multiplicity matrix-element evaluations still account for a non-negligible fraction of the total runtime in modern simulation workflows [7].

A broad range of strategies has been developed to address this challenge. On the one hand, significant speed-ups have been achieved through hardware acceleration. In particular, GPU based computation, pioneered with MadGraph4 [8] more than 10 years ago [9–11], are now production ready [12–15]. Complementary developments, such as the PEPPER framework [16], provide process specific GPU-accelerated event generation, including both fast amplitude evaluation and simplified integration routines.

In parallel, modern machine-learning techniques [17] have emerged as a powerful tool to accelerate event generation [18]. Considerable progress has been made in improving phase-space sampling through importance sampling, from early neural network approaches [19–21] to normalizing flows [22–27] and dedicated frameworks such as MADNIS [28–30]. While these methods can dramatically improve integration efficiency and unweighting, they typically leave the matrix-element evaluation itself untouched.

A more radical direction is to replace expensive first-principles amplitude calculations by fast machine-learning surrogates. Early work demonstrated the feasibility of learning scattering amplitudes or matrix-element weights directly from data using neural networks [31–36]. Since then, a wide range of increasingly structured approaches has been explored. These include combinations with multi-stage unweighting strategies [37–40], Lorentz-equivariant architectures [41–43], and the incorporation of learned uncertainty estimates [44–46]. An important development in this context is the use of physics-informed surrogate targets, in which neural networks learn ratios or correction factors with respect to analytically motivated approximations rather than absolute amplitudes or weights. In particular, the works of Ref. [34, 35] already exploit factorization properties of QCD amplitudes to construct surrogates with reduced dynamic range, significantly simplifying the learning task.

In this work, we pursue a closely related but distinct strategy that embeds QCD factorization more directly into the surrogate construction. Rather than correcting an analytic approximation at fixed multiplicity, we exploit the universal factorization structure of QCD amplitudes to relate an n-parton configuration to an exact reduced $(n-1)$-parton process. Our approach is rooted in the Catani–Seymour dipole formalism [47, 48], which provides an exact kinematic mapping between resolved and unresolved configurations. Starting from amplitudes for related processes with fewer final-state partons, we construct a factorized approximation of the full multi-jet amplitude and train a neural network to learn a correction factor that accounts for the residual difference to the exact result. This hybrid approach combines the strengths

of theory-driven factorization and data-driven learning. By construction, a large fraction of the kinematic complexity is absorbed into the reduced-multiplicity amplitude. Therefore, the neural network only needs to learn a smoother, lower-variance function that is free of explicit singularities. This leads to improved accuracy and robustness, particularly for high-multiplicity final states and in regions of phase space that are challenging for purely data-driven surrogates.

We apply this method to multi-jet production in association with a Z boson at leading order (LO) and show that factorization-based surrogates achieve high per-event accuracy across the full phase space. We further demonstrate how learned uncertainty estimates can be used to deploy the surrogate selectively within event generation, ensuring controlled numerical precision while achieving substantial speed-ups relative to standard matrix-element evaluations. Our approach complements advances in GPU acceleration and phase-space sampling and provides a viable path towards scalable event generation for high-multiplicity final states at future collider experiments as well as for the success of the high-luminosity runs [7].

The paper is organized as follows: In Sec. 2, we introduce the factorization ansatz underlying our surrogate construction, based on the Catani–Seymour factorization formalism. In Sec. 3, we describe the neural-network architecture and training strategy. In Sec. 4, we apply the method to Z+multi-jet production at leading order and assess the surrogate accuracy. In Sec. 5, we demonstrate its use within event generation and quantify the achievable speed-up. We conclude in Sec. 6.

2 Splitting kernel factorization

In this section, we introduce the factorization ansatz used for our neural-network surrogates. These formulas are based on the Catani–Seymour (CS) factorization, which accurately reproduces the expected behavior of QCD amplitudes in both the soft and collinear limits [47]. They ensure exact momentum conservation by redefining the momenta of the emitter and spectator particles, allowing the dipole terms to be applied even outside the strict soft or collinear regimes without violating momentum conservation or producing off-shell emitters. This section focuses on the conceptual aspects, with the full-fledged formulas provided in App. A.

2.1 Factorization ansatz

Given a scattering process with n final-state particles and at least one quark/gluon, we define the ensemble of reduced processes as the set of all $n-1$-particle final states obtained by absorbing one of the original quark/gluon into a pair of emitter and spectator particles. This is schematically expressed as:

$$(p_a, p_b \to p_1, \ldots, p_i, p_j, p_k, \ldots, p_n) \Longrightarrow \left\{(p_a, p_b \to p_1, \ldots, \tilde{p}_{ij}, \tilde{p}_k, \ldots, p_{n-1})\right\}_{(i,j,k)} \tag{1}$$

Here, p_i, p_j, p_k denote the four-momenta of the emitter, emitted parton (quark/gluon), and spectator, respectively. The momenta $\tilde{p}_{ij}, \tilde{p}_k$ are redefined according to the CS prescription (See App. A, Eq.(31)). The notation (i, j, k) runs over all valid combinations of emitter, emitted, and spectator particles.

We define the matrix-element squared, summed and/or averaged over colours and spins, of a process with n final-state particles as

$$A_n \equiv \langle |\mathcal{M}(p_a, p_b \to p_1, \ldots, p_n)|^2 \rangle \, . \tag{2}$$

which we will simply refer to as the *amplitude* in the following. Inspired by the CS formalism, we propose a general ansatz to approximate the amplitude of the full n-particle process, A_n,

in terms of a corresponding reduced $(n-1)$-particle amplitude, A_{n-1}. This factorized form is given by

$$A_n \approx A_{n-1} \cdot F^r_{ij,k} \,. \tag{3}$$

The reduced amplitude A_{n-1} depends on the set of reduced momenta, as shown in Eq.(1), for given particles i, j, k. The function $F^r_{ij,k}$ represents an approximate splitting kernel, which depends on the kinematics of particles i, j, k, as well as the nature of the emitted radiation r.

The factorization ansatz introduced in Eq.(3) is derived from the Catani–Seymour dipole formalism, subject to two simplifying assumptions. First, we neglect spin–helicity correlations by replacing the spin-dependent dipole kernel with its spin-averaged counterpart. Second, we adopt the leading-colour approximation, reducing the full colour-correlated structure to its dominant contribution. Under these assumptions, the dipole factorization takes a simplified scalar form, which serves as the foundation of our ansatz. In Appendix A, we present the Catani–Seymour dipole formalism for both initial- and final-state radiation with a final-state spectator, and we compare our factorization ansatz directly to the corresponding exact dipole expressions.

The choice of the spectator particle in the dipole construction is, in principle, arbitrary. For a fixed selection of emitter and emitted parton, different choices of the spectator lead to distinct reduced kinematics, due to the momentum redefinition applied to both the emitter and the spectator. In our study, we observed that restricting the choice of spectator to final-state jets improves the performance of the neural network, leading to better accuracy. Therefore, for both practical and methodological reasons, we restrict our analysis in this paper to cases where the spectator is chosen as the most energetic final-state jet, in the laboratory frame, that is neither the emitter nor the emitted parton.

2.2 Radiation type and rank

Usually, a process admits several reduced processes. Equation (3) requires only one reduced process, implying that we must choose both the type of radiation and the specific emitter–jet combination to use in our ansatz.

As the CS factorization accurately reproduces the QCD behavior in the soft and collinear limits, Eq.(3) provides higher accuracy for events near these singular regions. To achieve the best possible approximation, we rank all candidate reduced processes for a given event according to the degree of singularity of their corresponding radiation, and select the most singular ones.

For example, consider the process $d\bar{d} \to Z g_1 g_2 g_3 g_4$. If we look to the possible gluon splitting of the form $g_f \to g_f + g_f$, there are six distinct combinations where the emitted particles p_i, p_j are

$$\{g_i, g_j\}_{(i,j)=(1,2),(1,3),(1,4),(2,3),(2,4),(3,4)} \tag{4}$$

excluding symmetric configurations. By computing the scalar product $p_i\, p_j$ for each pair, we can rank the radiations from the most singular (smallest $p_i\, p_j$) to the least singular (largest $p_i\, p_j$). We define the rank of a radiation as its order in a scale of the most singular radiations of the same type. The first rank corresponds to the most singular radiations of the corresponding type, the second rank to the second most singular, and so on.

Additionally, a given process may contain several distinct radiation channels, determined by the identities of the emitter and emitted particles. Since each radiation type is associated with a different splitting kernel, we introduce the following radiation labels to distinguish them

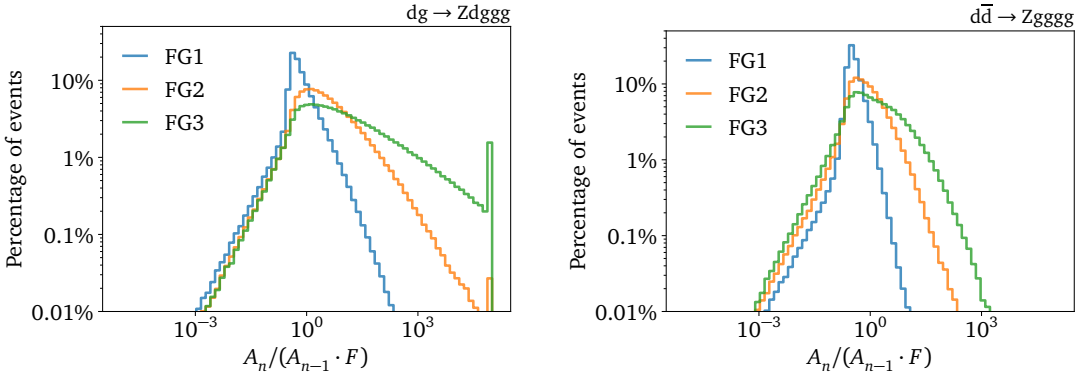

Figure 1: Approximation quality comparison between FG1, FG2, and FG3 radiations for dg → Zdggg (left) and for dd̄ → Zgggg (right), obtained from unweighted samples.

- FG1: Final state gluon radiation ($g_f \to g_f + g_f$);
- FQ1: Final state quark radiation ($q_f \to q_f + g_f$);
- IG1: Initial state gluon radiation ($g_i \to g_i + g_f$);
- IQ1: Initial state quark radiation ($q_i \to q_i + g_f$).

The presence of the index 1 inside the radiation labels is related to the radiation rank and can be replaced by another digit accordingly. This means that FG2 corresponds to the second rank radiation within the final state gluon radiation.

In Fig. 1, we illustrate the factorization quality (ratio between the target and the surrogate) of the first three ranks for $g_f \to g_f + g_f$ splittings for dg → Zdggg and dd̄ → Zgggg. Since, in our ansatz, the surrogate output will eventually be rescaled by a neural-network correction factor, the absolute value of the ratio is not particularly relevant. What matters instead is the shape of this distribution. A narrower and less dispersed distribution implies that the correction factor the network must learn is simpler and more stable, leading to a more efficient and accurate training procedure.

As expected, the most singular radiation yields the highest quality, while radiations with lower singularity exhibit broader distributions. Moreover, we notice that dd̄ → Zgggg have better approximations than dg → Zdggg. In fact, having 4 gluons in the final state, it has more radiations of the same type ($g_f \to g_f + g_f$) in respect to the other process, increasing the probability of finding a more singular pair, and so, a better factorization for the event. This behaviour confirms that the approximation quality is significantly higher for singular configurations (soft and/or collinear emissions) compared to the rest of the phase space.

In Fig. 2, we evaluate the performance of the factorization ansatz as a function of radiation type for the processes dg → Zdggg and dd̄ → Zgggg, including both initial- and final-state radiation. We find that final-state gluon radiation exhibits higher approximation accuracy in both cases. This trend can be attributed to combinatorial effects associated with the number of gluons, which increases the probability of identifying a softer or more collinear pair.

2.3 Double factorization

Using the CS formalism, we can iteratively absorb jets from the final state, thereby reducing the number of particles in the final configuration by one at each step. This procedure allows us to transition from a final state with n-particles to one with $(n-2)$-particles through successive

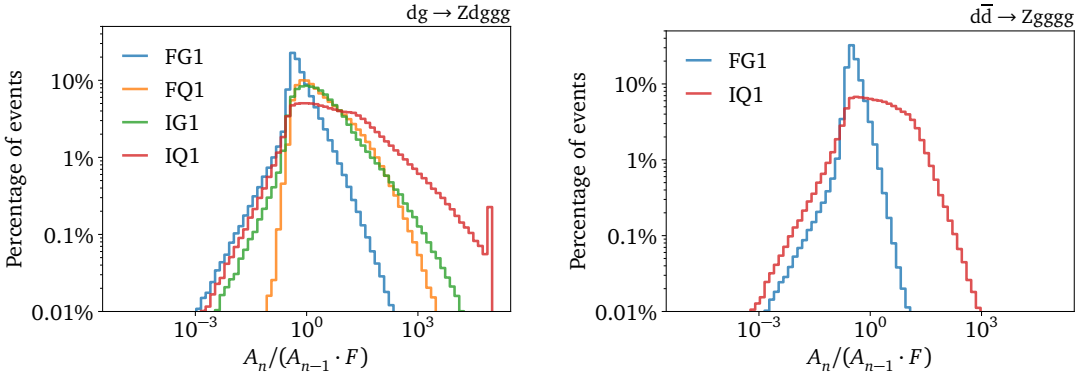

Figure 2: Approximation quality comparison between different radiation types: FG1, FQ1, IG1 and IQ1 radiations for dg → Zdggg (left), and FG1, and IQ1 radiations for $d\bar{d} \to Zgggg$ (right), obtained from unweighted samples.

reductions.

In the first step, one radiation is absorbed from the n-particle final state, resulting in a reduced process with $(n-1)$-particles. Then, a second radiation is absorbed to arrive at a $(n-2)$-particle state. The second reduction involves a splitting that depends on the reduced particle $\tilde{p}_{ij}$ produced during the first absorption. The approximated amplitude can thus be expressed as

$$A_n \approx A_{n-1} \cdot F^r_{ij,k} \approx A_{n-2} \cdot F^r_{\tilde{ij}l,k'} \cdot F^r_{ij,k}\,, \tag{5}$$

where $F^r_{\tilde{ij}l,k'}$ is the splitting kernel corresponding to the second reduction, and it depends on the momenta of the particles $\tilde{p}_{ij}$, p_l, and $p_{k'}$. The kinematics of the second reduction are therefore influenced by the first one, since one of the particles p_l or $p_{k'}$ may also serve as the spectator in the initial reduction.

In principle, one could select a radiation for the second reduction that does not involve the reduced particle $\tilde{p}_{ij}$ generated in the first step. However, in our study, we observed better performance from networks in which both reductions share the same reduced particle. For this reason, we restrict the radiation combinations to those that involve $\tilde{p}_{ij}$ in both steps. Similarly, although it is possible to mix different radiation types between the first and second reduction, we did not observe any significant benefit from doing so. Therefore, for simplicity, we limit our analysis to double factorizations constructed from identical radiation types.

In Fig. 3, we compare the approximation accuracy of the double factorization with that of the single factorization. To distinguish the notation, we denote the double factorization with a superscript square, for example FG1^2. Overall, the double approximation tends to exhibit reduced accuracy. While the central value of this approximation is of limited relevance – since it can be corrected by the neural network – it is worth noting that the double factorization estimate shows a noticeably broader spread. This implies that the neural network will need to handle a wider range of values, which may lead to a decrease in performance. Nevertheless, evaluating the double approximation is approximately ten times faster (see Tab. 1), which could make it useful for specific applications (see Section 5).

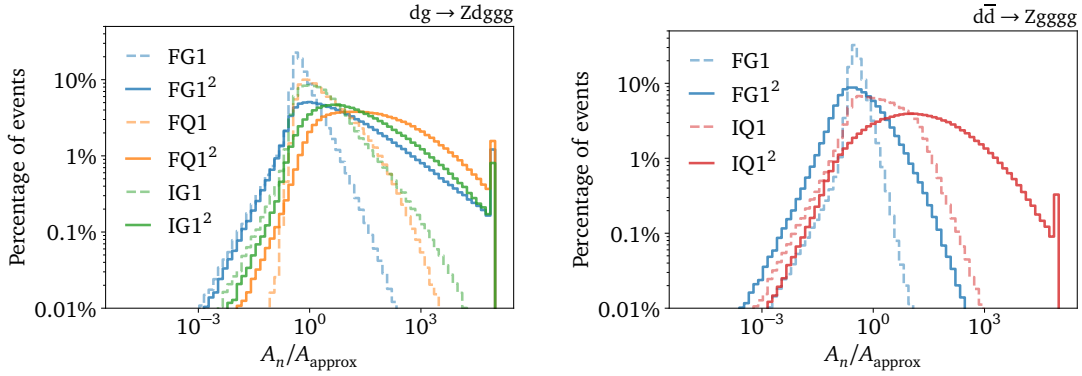

Figure 3: Approximation quality comparison between single and double factorizations for dg → Zdggg (left) and for $d\bar{d} \to$ Zgggg (right).

3 Machine-learning surrogate

Our aim is to develop a fast and accurate neural surrogate that leverages the factorized structure of the amplitude to approximate the full result. The factorization ansatz introduced in the previous section isolates the dominant singular behavior, leaving the network to learn a smooth correction factor with reduced variance and without large variations associated with on-shell resonances or explicit soft/collinear singularities.

All datasets used for training, validation, and testing in this work are generated with MG5aMC, using its default event-generation settings. Unless stated otherwise, the phase-space cuts are fixed to

$$p_{T,j} > 20 \text{ GeV}, \qquad \Delta R_{jj} > 0.4, \qquad |\eta_j| < 5\,. \tag{6}$$

These cuts define the reference phase-space region used throughout our analysis and correspond to a standard choice for LO event generation, needed to avoid infrared and collinear singularities. Importantly, no fine-tuning of the cuts was performed to make the ML-based surrogate more optimal or to artificially enhance its performance.

3.1 Preprocessing and heteroscedastic loss

Our training targets are exact theoretical amplitudes $A_{\text{true}}(x)$ for given phase-space points x, as defined in Eq.(2). Rather than regressing directly on the amplitude itself, which can span many

Coefficient	Evaluation time [s]
A_n	$1.3 \cdot 10^{-3}$
A_{n-1}	$1.0 \cdot 10^{-4}$
A_{n-2}	$1.2 \cdot 10^{-5}$
c_θ	$2.0 \cdot 10^{-8}$

Table 1: Summary of the average evaluation time per event for the process $d\bar{d} \to$ Zgggg and the corresponding reduced processes.

orders of magnitude across phase space, we train the network on a logarithmic representation,

$$\ell_{\text{true}}(x) = \frac{\log A_{\text{true}}(x) - \mu_{\text{train}}}{s_{\text{train}}}$$
$$\text{with} \quad \mu_{\text{train}} = \left\langle \log A_{\text{true}}(x) \right\rangle_{x\sim D_{\text{train}}} \qquad s^2_{\text{train}} = \left\langle (\log A_{\text{true}}(x) - \mu_{\text{train}})^2 \right\rangle_{x\sim D_{\text{train}}}, \tag{7}$$

including a linear standardization computed on the training data, where $\langle\cdot\rangle_{x\sim D_{\text{train}}}$ denotes an average over the training set. All network predictions are made in this standardized log-amplitude space and are mapped back to amplitude level only when evaluating performance or deploying the surrogate. This preprocessing stabilizes training and renders the learning problem numerically well-conditioned.

In this setup, there is no intrinsic measurement noise or stochasticity in the training data. Deviations between predictions and targets therefore arise purely from epistemic sources, reflecting our incomplete knowledge of the true amplitude function. These include (i) *finite-data and training-related effects* due to insufficient coverage of phase space and variability in the optimization procedure, which vanish in the limit of infinite data and perfect training, as well as (ii) a residual *model bias* originating from limitations of the factorization ansatz or the expressiveness of the neural network, which does not vanish even in the infinite-statistics limit.

We nevertheless employ a heteroscedastic-style objective to let the network predict, alongside its mean log-amplitude prediction $\ell_\theta(x)$, an input-dependent reliability proxy $\sigma_\theta(x)$, where θ denotes the trainable network weights. The loss can then be written as

$$\mathcal{L}_{\text{het}} = \left\langle \frac{(\ell_{\text{true}}(x) - \ell_\theta(x))^2}{2\sigma^2_\theta(x)} + \log \sigma_\theta(x) \right\rangle_{x\sim D_{\text{train}}}. \tag{8}$$

Compared to a standard mean-squared-error loss, this objective allows the network to dynamically downweight difficult regions of phase space during training. While Eq.(8) matches the algebraic form of a Gaussian likelihood, σ_θ is not interpreted as data (aleatoric) noise. Instead, it serves as a learned weighting term during optimization, with the following practical benefits:

1. Downweighting of rare or hard-to-learn configurations, leading to improved global accuracy and training stability.
2. Per-event confidence proxies that can be exploited in uncertainty-weighted combinations of multiple models, yielding more robust ensemble predictions.

Importantly, $\sigma_\theta(x)$ is *not* a measure of epistemic uncertainty in the statistical sense. Since it is learned jointly with the mean prediction $\ell_\theta(x)$, it absorbs a mixture of finite-data effects, optimization variability, and residual model bias.

More fundamentally, attempting to represent epistemic uncertainty by a point-wise quantity such as $\sigma_\theta(x)$ would destroy correlations between different phase-space points. Such correlations are essential for the consistent propagation of uncertainties to derived quantities, for example integrated or differential cross sections. To retain them, epistemic uncertainty must instead be extracted from *between-model* variability, for instance by using independently trained ensembles or Bayesian neural networks (BNNs) [49–54], which approximate sampling from the posterior $p(\theta|D_{\text{train}})$ and thus yield coherent, correlated uncertainty estimates across phase space.

We also tested BNNs, which are, in principle, well suited for estimating epistemic uncertainties. In practice, however, they produce central predictions comparable to those of our standard networks, while introducing increased training instability, higher model complexity, and slower inference. Moreover, when training a BNN with the heteroscedastic loss, the

$\sigma_\theta(x)$ term tends to absorb most epistemic variation, consistent with the expectation that these contributions become entangled in the zero-noise limit. A proper disentangling of epistemic components would require a more structured training procedure, such as the multi-step approach outlined in Ref. [55]. Since the focus of this work is on demonstrating a proof of concept for factorization-based amplitude surrogates, we leave such developments to future work.

Performance measure

For evaluation and physics validation, we quantify surrogate accuracy at the amplitude level through the relative deviation

$$\Delta(x) = \frac{A_{\text{true}}(x) - A_\theta(x)}{A_{\text{true}}(x)}\,, \tag{9}$$

which will be the primary performance metric used in Sec. 4. For small deviations, training in log-amplitude space is directly related to this quantity, since

$$\mathcal{L}_{\text{het}} \propto (\ell_\theta(x) - \ell_{\text{true}}(x))^2 = \frac{1}{s^2_{\text{train}}} \log^2(1-\Delta(x)) \simeq \frac{\Delta^2(x)}{s^2_{\text{train}}} + \mathcal{O}(\Delta^3)\,. \tag{10}$$

This means minimizing the mean-squared error in log space corresponds approximately to minimizing the squared relative deviation Δ^2, while avoiding numerical instabilities associated with large dynamic ranges in $A(x)$. The heteroscedastic loss therefore acts as a stabilized and uncertainty-weighted version of relative-error minimization.

3.2 Factorization network

Each factorization network uses a factorization approximation to reconstruct the full amplitude starting from a reduced one. Specifically, the network outputs a correction factor c_θ, which rescales the factorized approximation to match the true value. For a single factorization model, the predicted amplitude takes the form:

$$A_{n,\theta} = c_\theta \cdot A_{n-1} \cdot F^r_{ij,k} \tag{11}$$

where the radiation label r specifies the radiation type. For a double factorization model, the predicted amplitude is:

$$A_{n,\theta} = c_\theta \cdot A_{n-2} \cdot F^r_{\widetilde{ijl},k'} \cdot F^r_{ij,k} \tag{12}$$

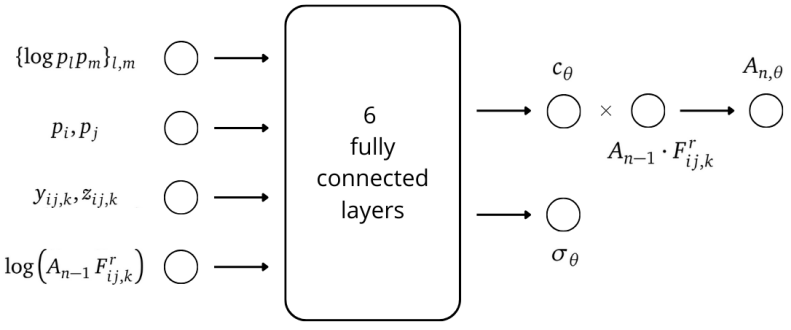

Figure 4: Representation of the single factorization neural network.

NN hyper-parameters	Value
Nodes per layer	[128, 256, 256, 128, 128, 64]
Total model parameters	$1.6 \cdot 10^5$
Activation function	Gelu
Optimizer	AdamW
Initial learning rate	10^{-3}
Final learning rate	10^{-8}
Learning rate scheduler	CosineAnnealing
Max epochs	1000
Batch size	256
Training events	$5 \cdot 10^5$
Validation events	$1 \cdot 10^5$

Table 2: Summary of the hyper-parameters used in a factorization neural network.

As a baseline, we also consider a *No Factorization* (NoFa) model, which predicts the amplitude directly without using any analytical factorization structure. In this case,

$$A_{n,\theta} = c_\theta \tag{13}$$

where c_θ is obtained by de-processing the raw network output c'_θ using the same transformation applied to the training target amplitudes $c_\theta = \exp\left(c'_\theta \cdot s_{\text{train}} + \mu_{\text{train}}\right)$, allowing the network to effectively cover several orders of magnitude.

In our study, we also explored a multi-radiation factorization model, in which a single network predicts the full amplitude as a linear combination of single-radiation approximations. The ansatz for this model is

$$A_{n,\theta} = \sum_r^{N_r} c_\theta^r \cdot A_n^r \cdot F_{ij,k}^r \tag{14}$$

where the sum over r includes the N_r selected radiations. The reduced amplitude A_n^r differs for each radiation and must be computed N_r times for each event. In our tests, this approach did not show a clear improvement in accuracy compared to the single-radiation model. Given its higher computational cost, we chose not to include this model in the final results.

The technical details of our neural-network architecture are summarized in two tables. Table 2 lists the hyperparameters used for all networks in this work, while Table 3 specifies the input features for a single-factorization network. The kinematic log-invariants and the logarithmic factorization inputs are also standardised, in order to account for their disparate orders of magnitude. To guide the model toward the most relevant configurations, we reorder the momenta of all possible emitter and emitted particles for the corresponding radiation type

Input feature	Variable
Kinematic log-invariants	$\{\log p_l p_m\}_{l,m}$
Emitter and emitted four-momenta	p_i, p_j
Radiation variables	$y_{ij,k}, z_{ij,k}$
Log-factorization factors	$\log\left(A_{n-1} F_{ij,k}^r\right)$

Table 3: Summary of input features used in a single factorization model.

before computing the kinematic invariants. This rearrangement is based on the singularity associated with each radiation and encodes the radiation rank into the particle ordering.

For the NoFa model, no reordering is applied, and no factorization information is provided as input. In addition, for the double-factorization model, two additional outputs are included to provide to the network the extra radiation variables.

Network ensembling

Ensemble neural network models combine the predictions of multiple individual networks to improve robustness and accuracy. By aggregating diverse models, ensemble methods reduce over-fitting and provide a broader, more reliable representation of the underlying physical processes. In this work, we employ an ensemble of simple neural networks, each based on a distinct factorization approximation. It differs from the multi-radiation single network, where the same network predicts all the correction factors of the ansatz.

We define our ensemble model as a linear combination of factorization models. Given a set of factorization models $\{1,\dots,n_{\text{mod}}\}$, a model j predicts for a given event i an output $(A^i_{\theta_j}, \sigma^i_{\theta_j})$, where $A^i_{\theta_j}$ is the predicted amplitude for the event i and $\sigma^i_{\theta_j}$ is the corresponding predicted uncertainty. The predicted ensemble amplitude is given by

$$A^i_{\theta_e} = \sum_{j=1}^{n_{\text{mod}}} w^i_j \cdot A^i_{\theta_j} \,, \tag{15}$$

where the weights w^i_j are computed from the inverse of the covariance matrix C_i^{-1} of the model predictions [56, 57]

$$w^i_j = \frac{\sum_k (C_i^{-1})_{jk}}{\sum_{j,k} (C_i^{-1})_{jk}} \,. \tag{16}$$

To estimate C_i, we first determine the correlation matrix ρ between two different models. The correlation ρ_{ab} is defined as the average correlation between the per-point accuracies of models a and b over the entire validation dataset. This correlation is treated as a global quantity and kept fixed across phase space. We also investigated a per-point, phase-space–dependent correlation, but found neither improvement nor degradation in performance. Defining ρ_{ab} in terms of accuracy allows us to quantify whether two models tend to make similar prediction errors. The covariance between models a and b for event i is then given by

$$C_{i_{ab}} = \rho_{ab} \cdot \sigma^i_{\theta_b} \cdot \sigma^i_{\theta_b} \tag{17}$$

from which we obtain the inverse matrix C_i^{-1}. Finally, the predicted uncertainty for the ensemble model is

$$\sigma^i_{\theta_e} = \frac{1}{\sqrt{\sum_{j,k} (C_i^{-1})_{jk}}} \,. \tag{18}$$

4 Application to Z with multi-jet production

As previously noted, we benchmark our CS-based ansatz using Z+jet production. This choice is mainly motivated by its use in previous surrogate studies [34, 42], which facilitates direct comparison with existing approaches. Nonetheless, our method will exhibit even greater advantages for Beyond Standard Model processes accompanied by jets, or more generally for scenarios involving multiple resonances and/or scales in conjunction with jets. Within the Standard Model, particularly suitable candidates include vector-boson fusion with jets or fully decayed top-quark production with jets. A comprehensive treatment of these processes, together with full automation of the proposed methodology, is deferred to future work.

4.1 Performance for quark-gluon initial state

We begin by comparing single factorization models applied to different types and ranks of radiation within the same process. Specifically, we consider the process dg → Zdggg, and evaluate the following radiation configurations: FG1, FQ1, IG1, IQ1, for the radiation types, and FG1, FG2, FG3, for the radiation ranks.

Figure 5 compares the performance of the models introduced above. First, we observe that all radiation-based models outperform the NoFa model, confirming the benefit of incorporating the factorization structure. In the left panel, the factorization models achieve similar average accuracy despite originating from different underlying factorization qualities. The two bests models are here FG1 and IG1 with very similar mean accuracy but quite different shape: IG1 being more peak (i.e. consistent with the average) while FG1 has a broader shape, with especially more event with better precision than 10^{-3}. Both models have quite similar rate of events with low accuracy ($> 10^{-2}$).

The picture is simpler in the right panel, where we present the accuracy as a function of the approximation rank. Selecting the most singular approximation yields a clear improvement in accuracy, both in terms of the mean and the overall distribution. In contrast, there is no noticeable difference between choosing the second or third rank, indicating that the critical information lies primarily in the first-rank emission. This provides an a posteriori indication that using only one reduction step – the most singular one – is a reasonable choice.

We now compare the performance of the double factorization models across different ra-

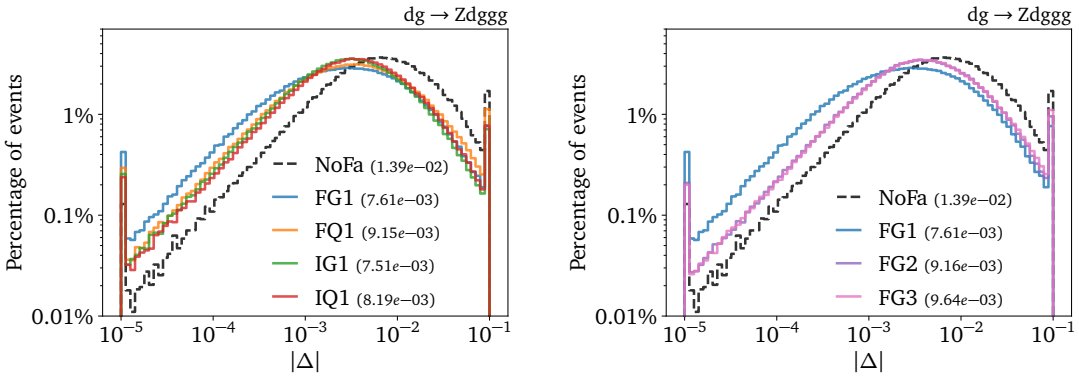

Figure 5: Single factorization model accuracy comparison between different radiation types (left), and different radiation ranks (right) for the process dg → Zdggg. The values in parentheses are the mean accuracies over the whole test dataset.

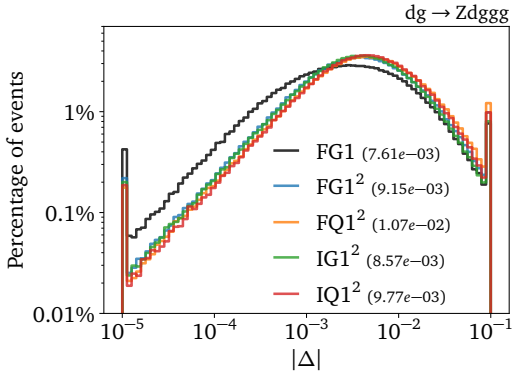

Figure 6: Double factorization model accuracy comparison between different radiation types for the process dg → Zdggg. The values in parentheses are the mean accuracies over the whole test dataset.

diation types: $FG1^2$, $FQ1^2$, $IG1^2$ and $IQ1^2$. In Fig. 6, we report the corresponding accuracy distributions. The double factorization models exhibits reduced accuracy compared to their single-factorization counterparts. This was obviously expected since the underlying approximation was much less precise to start with. Additionally, one can also see a much weaker dependence in the kernel used, which is also not very surprising given the distribution to learn by the neural network (Fig. 3) which shows that all curves are indeed quite similar. It is noteworthy that, for FG radiation, the double-factorization model $FG1^2$ exhibits performance that is very close to that of the single-factorization model of second rank, FG2. Finally, when comparing the double-factorization model to the NoFa baseline for this process, the improvement remains modest.

Ensemble factorization models

After evaluating the performance of individual radiation models, we now investigate how their predictions can be combined through network ensembles to improve accuracy and robustness. As previously discussed, each radiation model captures different aspects of the same underlying process, and even models with similar accuracy may encode different information due to the type and rank of the radiation. While this motivates the use of ensemble methods to leverage the diversity among models, a large part of that information is redundant and one need to be careful about the correlations between the various predictions.

As described before, we define an network ensemble as a combination of independently trained networks. In this context, how the individual predictions are recombined is crucial. A naive recombination that is independent of the predicted uncertainty can lead to suboptimal results. In our case, we observed that the uncertainty-weighted recombination described in Eq.(15) consistently outperforms any of the individual single models. From Fig. 7, we confirm that each ensemble achieves a better mean accuracy than its constituent individual models.

Interestingly, the overall accuracy of the ensemble does not depend solely on the accuracies of the individual models, but also on the radiation types and ranks used to construct the ensemble. Combining the most accurate individual models does not always yield the best ensemble performance, as shown by the ensemble [FG1, FQ1] achieving a better accuracy than [FG1, IG1], despite IG1 having better performance. This behavior highlights that networks trained on different radiations tend to extract complementary information from the same dataset, and that low error correlation, or model diversity, is a key factor for achieving strong ensemble re-

sults. We also experimented with modifying the loss and the learning algorithm to encourage such complementarity and improve accuracy, but without significant success.

On the right panel, we compare ensemble models of double factorization combined with single factorization one. These ensembles, even if yielding a lower accuracy due to the approximation used, will actually have better speed-up factor due to faster evaluation time.

Similarly, we observe that ensembles composed of three models outperform those composed of only two, although the performance gain is less significant than the improvement obtained when moving from a single model to an ensemble of two.

4.2 Performance for quark-quark initial state

So far, we have evaluated the performance of single and ensemble models on the process $dg \to Zdggg$ to investigate how the accuracy depends on the chosen factorization model. In this section, we extend the study to a different process: $d\bar{d} \to Zgggg$ with the main difference compare to the previous section is that all the gluon are in the final state.

Single and double factorization models

We begin by examining the performance of the individual factorization models. The left panel of Fig. 8 shows the single-factorization results. Comparing this plot with the corresponding one for the previous process (Fig. 5), we observe that the NoFa model achieves nearly identical accuracy in both cases. This is not the case for the factorization models: in particular, the best-performing model, FG1, attains substantially higher accuracy. This improvement appears to result from having all gluons—which dominate the radiation pattern—in the final state, thereby enhancing the effectiveness of the factorization ansatz.

The right panel of Fig. 8 shows the accuracy of the double-factorization model. Here as well, the performance remains significantly better than the NoFa baseline. Moreover, similarly to the $dg \to Zdggg$ case, we observe that the accuracy achieved by $FG1^2$ is nearly identical to that of the single-factorization FG2 model.

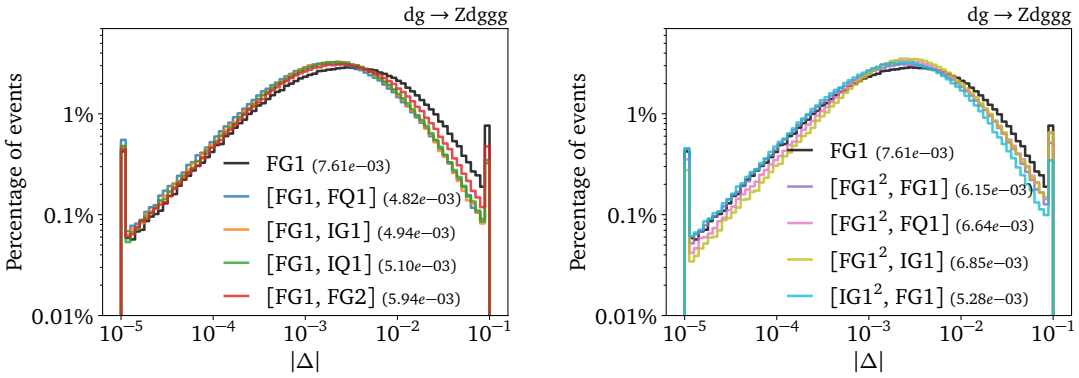

Figure 7: Ensemble factorization model accuracy comparison between single factorization (left), and double and single factorization (right) for the process $dg \to Zdggg$. The values in parentheses are the mean accuracies over the whole test dataset.

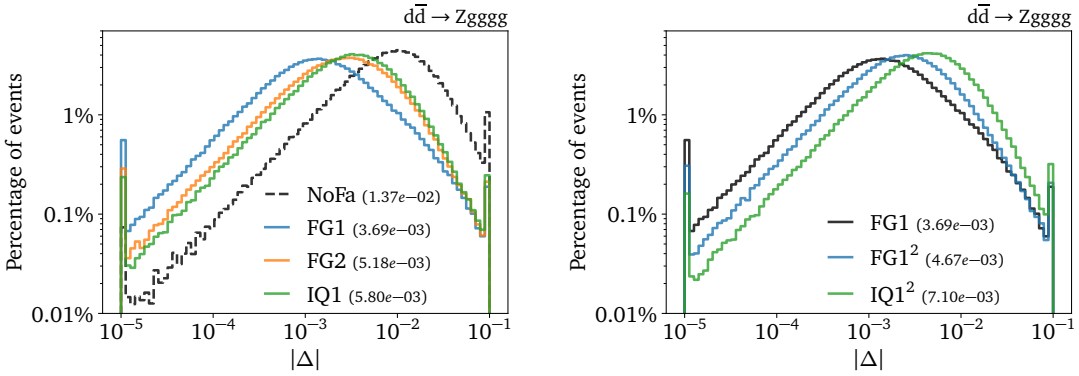

Figure 8: Accuracies (left) and cumulative accuracy (right) of single radiation models on $d\bar{d} \to Zgggg$ process. On the left, the values in parentheses are the mean of the distribution.

Ensemble factorization models

Finally, we examine the accuracy of the ensemble models for the process $d\bar{d} \to Zgggg$. As in the previous process, we confirm that the ensemble models outperform the individual ones across the entire test dataset. In the left panel of Fig. 9, we directly compare each ensemble with its best corresponding individual model. We observe that the ensembles reduce the large-error tail more significantly than the small-error tail, indicating that combining different models is especially beneficial for events that are otherwise difficult to predict accurately.

The right panel of Fig. 9 shows ensembles that combine double- and single-factorization models. In this case, the best-performing ensemble is [IG1^2, FG1]. This result suggests that the most effective strategy is not to combine the two most accurate models, but rather those that provide complementary information. Furthermore, this ensemble achieves an accuracy comparable to the best ensembles composed exclusively of single-factorization models, while offering faster evaluation times.

This process corresponds to the one used in [42], which employs a Lorentz-equivariant network. A direct comparison of the achieved accuracy is therefore possible. In that work,

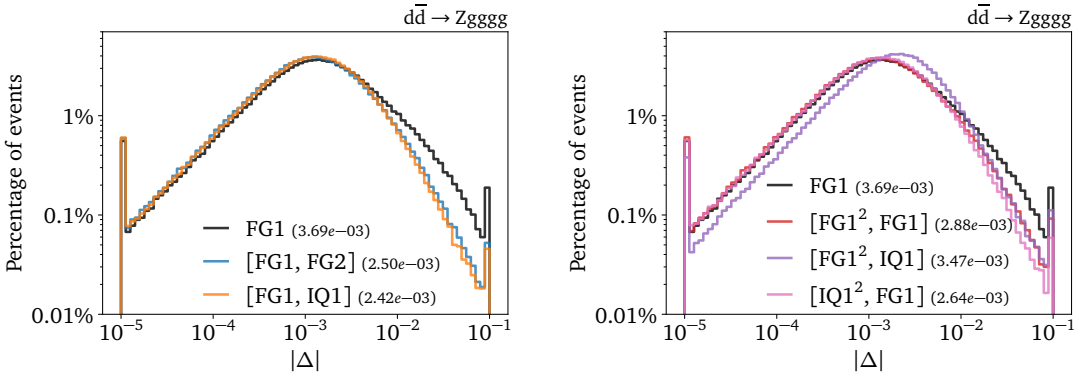

Figure 9: Ensemble factorization model accuracy comparison between single factorization (left), and double and single factorization (right) for the process $d\bar{d} \to Zgggg$. The values in parentheses are the mean accuracies over the whole test dataset.

Neural network model	Mse	mean($\lvert\Delta\rvert$)
FG1	$1.5 \cdot 10^{-5}$	$3.69 \cdot 10^{-3}$
FG1^2	$1.8 \cdot 10^{-5}$	$4.67 \cdot 10^{-3}$
[FG1, FG1^2]	$8.7 \cdot 10^{-6}$	$2.88 \cdot 10^{-3}$
L-GATr*	$\sim 8.4 \cdot 10^{-6}$	
[FG1, FG2]	$5.9 \cdot 10^{-6}$	$2.50 \cdot 10^{-3}$
[FG1, IQ1]	$5.5 \cdot 10^{-6}$	$2.42 \cdot 10^{-3}$

Table 4: Comparison of the various network models in terms of logarithmically standardized mean squared error (mse)on $d\bar{d} \to Zgggg$ process, including L-GATr [42](trained on $4 \cdot 10^5$ samples). The last column reports, for our models, the mean accuracy metric used in this paper. Model are ordered from less precise to the most accurate one (according to the mse).

accuracy is not reported directly; instead, the authors use the mean squared error (MSE) on the logarithmic standardized amplitude as a measure of network performance. For comparison purposes, we report in Table 4 the corresponding values for several of our models. These results indicate that the achieved accuracy is in the same ballpark as that of the L-GATr network, which remains more performant under this metric than any of our single-factorization models. However, our most accurate ensemble network surpasses it.

5 Achievable speed-up without double-unweighting

Having a fast and accurate surrogate is only useful if we can use is it practically. While we here focus on how to make LO phase-space integration faster, the algorithm that we will describe in this section can generally be applied to any computation relying on evaluating amplitudes. In the literature [37, 39, 40] a common approach when using surrogates is to use it within importance sampling, while relying on standard hit and miss to generate event following the surrogate density function, which is conceptually equivalent to other importance sampling algorithm (based on ML or not) [20, 22, 23, 26–30, 58–66]. In those methods, even in the perfect scenario, one need to evaluate at least one amplitude (and typically much more) for each unweighted event produced.

In this paper, we propose using the surrogate directly in place of the full computation when its estimated error is sufficiently small. This approach, described in more detail below, can significantly reduce the number of evaluations of the full amplitude—potentially eliminating them entirely at the generation stage. The main drawback is that it introduces an additional source of numerical uncertainty associated with the precision of the surrogate itself.

An often-cited concern with such a strategy is that the tails of distributions might be severely misrepresented. In practice, however, most observables $\mathcal{O}$ correspond to cross sections integrated over restricted regions of phase space. For a binned (differential) observable defined by a phase-space region $\Omega \subset \Phi$, for instance corresponding to a bin in a low-p_T tail, we can write

$$\mathcal{O} \equiv \int_\Omega \mathrm{d}x \, \frac{\mathrm{d}\sigma}{\mathrm{d}x} \approx \frac{1}{N} \sum_{i \in \Omega} w_i \,, \tag{19}$$

where the sum runs over all Monte-Carlo events that fall within the phase-space region Ω, and the weights w_i denote the usual event weights, including the amplitude, the parton distribution

functions, and the Jacobian associated with the phase-space measure. In the ideal case, the errors on the individual weights w_i are fully uncorrelated and thus

$$(\Delta\mathcal{O})^2 = \frac{1}{N^2}\sum_i (\Delta w_i)^2 \,. \tag{20}$$

In this case, the relative uncertainty on any observable $\mathcal{O}$ vanishes in the infinite–statistics limit

$$\frac{(\Delta\mathcal{O})^2}{\mathcal{O}^2} = \frac{\sum_i (\Delta w_i)^2}{(\sum_i w_i)^2} = \frac{\sum_i \left(\frac{\Delta w_i}{w_i}\right)^2 w_i^2}{\left(\sum_i w_i\right)^2} \leq \max_i\left(\frac{(\Delta w_i)^2}{w_i^2}\right)\frac{\sum w_i^2}{\left(\sum_i w_i\right)^2} \tag{21}$$

$$\Rightarrow \qquad \frac{\Delta\mathcal{O}}{\mathcal{O}} \leq \frac{1}{\sqrt{\alpha N}}\max_i\left(\frac{\Delta w_i}{w_i}\right), \tag{22}$$

where we have introduced the Kish effective sample size [67]

$$N_{\text{eff}} = \frac{\left(\sum_i w_i\right)^2}{\sum_i w_i^2} = \alpha N \qquad \text{with} \qquad \alpha \in [0,1]\,. \tag{23}$$

However, assuming full un-correlation is first unrealistic but also not fully general since one can think that for some part of the phase-space and/or for specific observable those errors will be fully correlated, i.e.

$$\Delta\mathcal{O} = \frac{1}{N}\sum_i \Delta w_i \,. \tag{24}$$

In that case, the relative uncertainty does not vanish anymore but is bounded as

$$\frac{\Delta\mathcal{O}}{\mathcal{O}} = \frac{\sum_i \Delta w_i}{\sum_i w_i} =\leq \frac{\sum_i w_i \,\max_i(\frac{\Delta w_i}{w_i})}{\sum_i w_i} = \max_i\left(\frac{\Delta w_i}{w_i}\right). \tag{25}$$

This means that even in the worst-case scenario, where all weights are biased in the same direction by, say, X%, all observables would be biased. But, the relative uncertainty on all observables would remain controlled and would correspond exactly to X%.

Consequently, the task at hand is to control the uncertainty of the surrogate across the full fiducial phase space. Before introducing our algorithm, we first review other sources of theoretical uncertainty to determine the required accuracy of the surrogate. For leading-order event generation, the dominant source of uncertainty arises from scale variation, which is typically of the order of fifty percent for the normalization and at least ten percent for the shape. A second source of uncertainty comes from the parton distribution functions, corresponding to an error of about three percent (and likely larger in the tails of distributions). Therefore, any numerical error below these scales across the full phase space can be considered acceptable. As a standard benchmark for this paper, we target a one-sigma error (66% confidence level) of approximately one percent.

As a matter of fact, none of our surrogates achieve this level of precision over the full phase-space. Therefore, we propose a mixed approach: the surrogate prediction for an event is going to be used only if its estimated relative predicted uncertainty, defined as $\sigma_{\text{nn}}/A_{\text{nn}}$, is smaller than a chosen uncertainty threshold U_{thr}. Events that do not satisfy this requirement are instead evaluated using the exact amplitude, ensuring accuracy over the full phase-space.

In order to be effective, the method deeply relies on the fact that our estimator of σ_θ behaves correctly. To assess that, we display in Fig. 10 and 11 (left panel) the correlation

between the estimated error from the surrogate and the actual error. On one hand, we see a clear tendency for σ_θ to be over-conservative, especially for very accurate prediction. On the other, we see a quite good correlation between the two quantities. However, one still needs to be careful since some under-estimation can still occur (with normal distribution and therefore for rare event). This is why we only use σ_θ as a proxy of the accuracy and not as the real measure. To quantify the additional uncertainty introduced by our approach, we introduce a tolerance parameter $\tau_{(x)}$, defined as the fraction of events in the final sample whose accuracy is worse than x%, which means $|\Delta| > x$. The tolerance is therefore

$$\tau_{(x)} = \epsilon_U \cdot f_U^{(x)}, \tag{26}$$

where ϵ_U is the fraction of events where $\sigma_\theta / A_\theta < U_{\text{thr}}$, and $f_U^{(x)}$ is the fraction of those events with accuracy larger than x.

While the choice of acceptable tolerance can vary from one application to another, we have pick for the case of LO phase-space integration a value of $\tau_{(3\%)} = 10^{-3}$. which, if we assume that everything is Gaussian (which is kind of reasonable) means a 1 sigma error of around 1 percent, our target for LO phase-space generation. Additionally, one should remember that this 1 sigma error at one percent for any observable is reached in the case where all the error are correlated, which is a very conservative upper-bound. This is confirmed by Fig. 10 and 11 (right panel) where we plot independently the events who over-estimate the amplitude from those who under-estimated. For those plots, we have selected only the events passing the respective thresholds, $U_{\text{thr}} = 0.0102$ and $U_{\text{thr}} = 0.0111$, such that we have $\tau_{(3\%)} = 10^{-3}$. Both curves are almost perfectly aligned, suggesting that the error behaves more like in the uncorrelated case than in the correlated one.

Once a tolerance is selected for a given application, we determine—via a simple scan—the optimal value of U_{thr} required to achieve that tolerance. We then compute the speed-up factor

$$f \equiv \frac{t_{\text{MG}}}{t_{\text{tot}}} = \frac{t_{\text{MG}}}{t_{\text{surr}} + (1 - \epsilon_U) \cdot t_{\text{MG}}}, \tag{27}$$

where t_{surr} is the time to evaluate the surrogate and t_{MG} is the time to evaluate the true amplitude within MG5aMC.

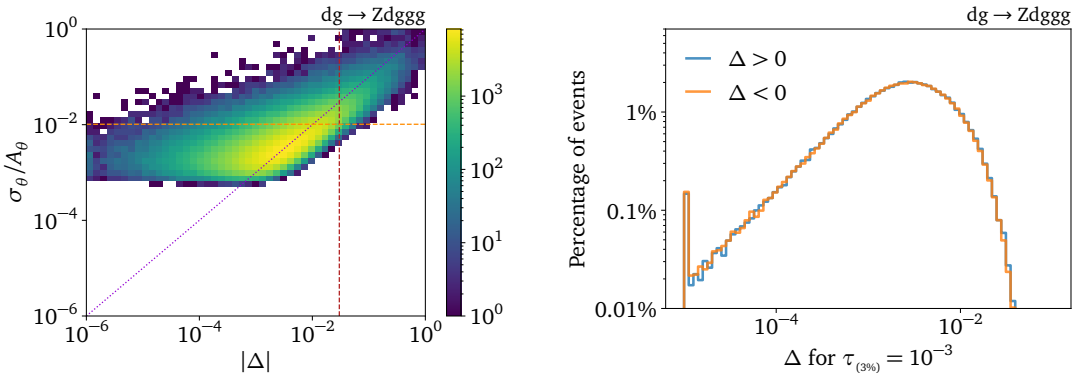

Figure 10: Left: Absolute accuracy vs. relative predicted uncertainty for IG1-model on the whole test dataset. The horizontal dashed line corresponds to U_{thr} for $\tau_{(3\%)} = 10^{-3}$. The vertical dashed line corresponds to the target accuracy $\Delta = 3\%$. Right: Accuracy distribution for the events using the surrogates for $\tau_{(3\%)} = 10^{-3}$ split into negative and positive contribution, for IG1-model on the dg → Zdggg process.

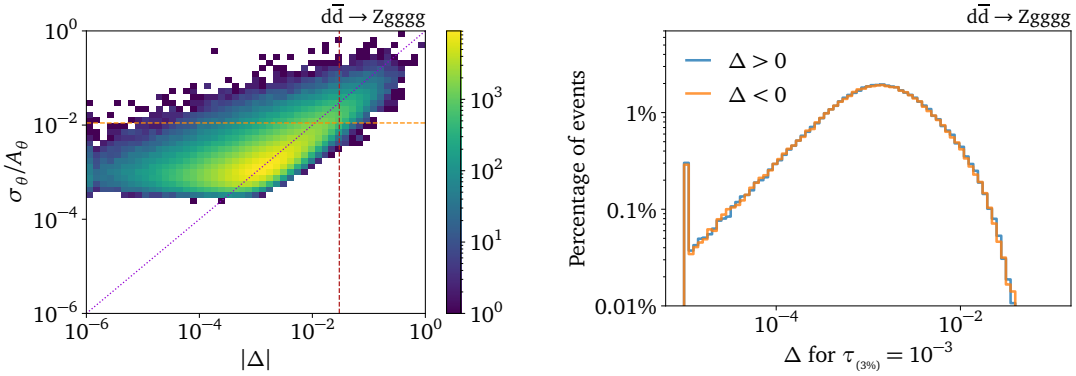

Figure 11: Left: Absolute accuracy vs. relative predicted uncertainty for FG1-model on the whole test dataset. The horizontal dashed line corresponds to U_{thr} for $\tau_{(3\%)} = 10^{-3}$. The vertical dashed line corresponds to the target accuracy $\Delta = 3\%$. Right: Accuracy distribution for the events using the surrogates for $\tau_{(3\%)} = 10^{-3}$ split into negative and positive contribution, for FG1-model on the $d\bar{d} \to Zgggg$ process.

As usual when quoting any effective gain factor, one need to stress underline hypothesis and context. The first hypothesis here is that the timing of the computation is highly dominating by the times needed to evaluate such amplitude, allowing us to not take into account the time needed to train such network efficiently, which is likely only valid for the massive HL-LHC simulation of CMS and ATLAS. The second one is that the application under consideration is highly dominated by the time needed to evaluate the amplitude, otherwise one need to apply Amdahl's law [68] to rescale the speed-up factor down.

This algorithm can naturally be extended to test multiple surrogates sequentially, leveraging the fact that some networks predict certain regions of phase space more accurately than others. For efficiency, we order the surrogates by evaluation cost, and both the speed-up factor and tolerance definition are updated accordingly. Combining two or more surrogates offers an additional advantage: the ensemble can itself serve as a surrogate. In practice, we first evaluate the faster surrogate; if predicted uncertainty is above our threshold, we then evaluate the second surrogate. The final decision is then based on the ensemble prediction rather than solely on the second surrogate, since the ensemble is at least as accurate and both components have already been computed, the additional cost is virtually zero.

Moreover, using two surrogates allows us to fine-tune the uncertainty thresholds, assigning different thresholds to each step. In our study, we found that choosing a smaller threshold for the first surrogate and a slightly larger threshold for the second surrogate (relative to a uniform U_{thr}) improves performance while maintaining the same tolerance. This approach lets us impose stricter constraints on the less accurate network and more relaxed constraints on the more reliable one.

Speed-up for dg → Zdggg

In Fig. 12, we present, in function of the internal threshold U_T, both the value of the tolerance for various target accuracy (left panel) and the achieved speed-up (right panel) for the process dg → Zdggg. For those particular plots, we choose to use the best performing model for this process, but other factorization ansatz have similar behaviour. On the left panel, the interesting point to note for our benchmark point, corresponding to $\tau_{(3\%)} = 10^{-3}$ (the green dashed line) is the value of the tolerance for the other threshold. We see that $\tau_{(1\%)} \sim 7 \cdot 10^{-2}$ which is for the 1

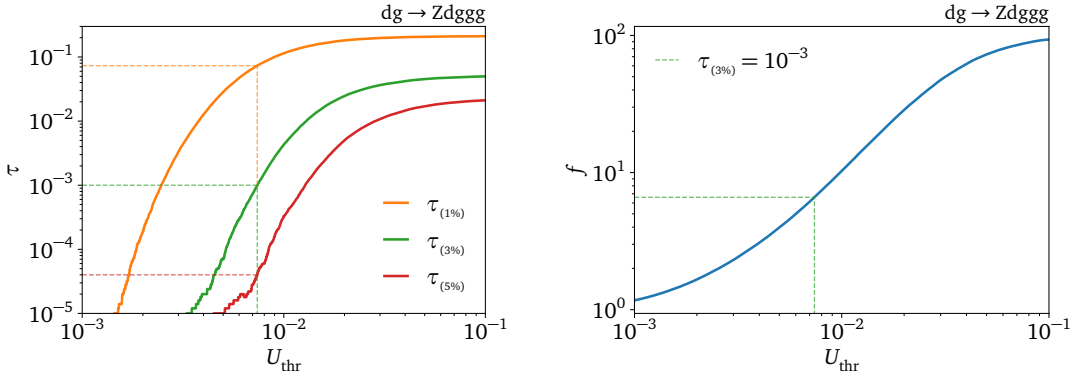

Figure 12: Tolerances (left) and Speed-up factor (right) vs. the uncertainty threshold U_{thr} for the ensemble [IG1^2, FG1] factorization model for the process dg → Zdggg. The plots are obtained using different uncertainty thresholds for the first and second surrogate, such that: $U_1 = U_{\text{thr}}$ and $U_2 = 1.4 \cdot U_{\text{thr}}$

percent is actually much better than one would expect given a pure normal distribution of error, which we relate to the fact that our network tend to be over-conservative. The $\tau_{(5\%)} \sim 4 \cdot 10^{-5}$ being more consistent with the theoretical expectation. This comforts us on the decision to use $\tau_{(3\%)} = 10^{-3}$ as a reasonable threshold. We will therefore only focus on $\tau_{3\%}$ afterwards.

In the right panel, we show the behaviour of the achieved speed-up. On the left side of the plot we find the region with stringent tolerance constraints, which leads to the rejection of most surrogate-evaluated events. As a consequence, the full amplitude must be computed for the majority of events, and the speed-up remains close to 1. Conversely, the right side of the plot corresponds to the region with looser constraints, where most events are accepted during the first surrogate evaluation. In this regime, the speed-up reaches its maximum value, approaching the ratio between the cost of evaluating the doubly reduced amplitude and that of the full one.

In Figure 13, we present the final speed-up factor for a given value of the tolerance at 3%, enlightening the value of $\tau_{(3\%)} = 10^{-3}$ as our standard benchmark point. In this plot, we compare the speed-up factors for our different surrogate networks. The results are split into two panels: the left shows single-network configurations, while the right shows ensembles of two surrogate networks (which therefore require two training). It can be seen that, although the no-factorization network is the fastest to evaluate, its effective speed-up factor is quite small due to its limited accuracy—both in predicting the amplitude and in estimating σ_{NN}). At the same time, the double-factorization model slightly outperforms the single-factorization model, primarily because it is approximately ten times faster to evaluate while still providing reasonably accurate results. Interestingly, this advantage disappears for stricter tolerances ($\tau_{(3\%)} \leq 3 \cdot 10^{-4}$).

Looking now at the result when using two different surrogates, one can observe different effect due to the subtle interplay between the speed and the accuracy of the surrogates. The winning strategy is actually to use the most accurate ensemble of double and single factorization, as shown in Fig. 7. Comparing the winning ensemble to the similar models, we find that this network gains both from having the most accurate first surrogate (IG1^2), and from the most accurate ensemble ([IG1^2, FG1]), and not just the most accurate second surrogate (IG1). We can also build ensemble that use NoFa as a first surrogate. However, the performances of those models are worse compared to ensembles of double and single factorization, thanks to their higher accuracy and more meaningful ensemble prediction.

As can be read on the graph, we reach for our benchmark point a speed-up factor of 6.5, that can be larger if one allows themselves to be less conservative than us on the allowed threshold. On the other direction, if one wants to be even more conservative of either $\tau_{(3\%)} = 10^{-4}$ or $\tau_{(3\%)} = 10^{-5}$, one still get speed-up of roughly 4 and 3 respectively.

Speed-up for $d\bar{d} \to Zgggg$

Finally, we present the results for the second process, that is $d\bar{d} \to Zgggg$. In Fig. 14, we show the tolerance (left panel) and the achieved speed-up (right panel) as functions of the internal threshold U_{thr}. Comparing the tolerance panel with that of the previous process (Fig. 12), we observe a similar behaviour. However, in this case the threshold U_{thr} corresponding to a given target tolerance is noticeably larger ($U_{\text{thr}} = 0.0084$ versus 0.0073), reflecting the higher accuracy of the surrogate, which permits looser constraints.

For the same reason, in the right panel, we see that the speed-up increases more rapidly as U_{thr} grows. Nevertheless, both processes reach the same maximum speed-up, since the evaluation times are identical; the difference lies in the tolerance values at which this plateau is achieved.

In Fig. 15, we present the final speed-up factor has a function of the tolerance (at 3 percent), for both individual (left panel) and ensemble networks (right panel). A clear improvement is observed when transitioning from the NoFa model to a single-factorization surrogate, and subsequently to a double-factorization surrogate. In this case, the double-factorization model is sufficiently accurate to provide nearly a factor-of-two improvement in speed-up compared to the single-factorization model.

For the ensemble models, we find that they achieve similar speed-up factors. As in the other process, the optimal strategy is to use the most accurate double-factorization network as the first surrogate. However, unlike the previous case, we do not observe significant differences between the most accurate ensemble model ([$FG1^2$, IQ1]) and the ensemble combining two highly accurate surrogates ([$FG1^2$, FG1]). For our benchmark, we obtain a speed-up factor of approximately 20. As for the previous process, this factor can be higher (and can reach 30) if one allows himself to be less conservative (that speed-up is reached for a still reasonable $\tau_{(3\%)} = 3 \cdot 10^{-3}$). On the opposite direction, the more accurate network allows to still achieve a sizable speed-up of 7, even in the very conservative limit ($\tau_{(3\%)} = 10^{-5}$).

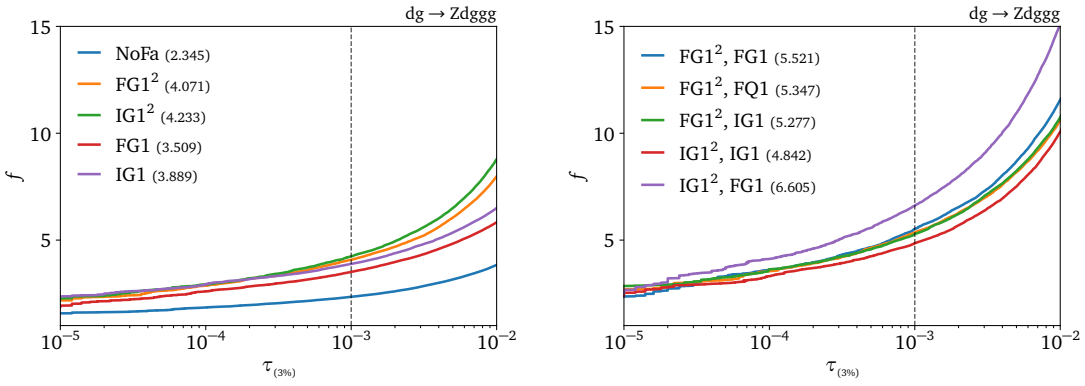

Figure 13: Speed-up factor comparison between different individual factorization models (left), and ensemble factorization models (right) for the process $dg \to Zdggg$. The values in parentheses are the speed-up factors corresponding for $\tau_{(3\%)} = 10^{-3}$.

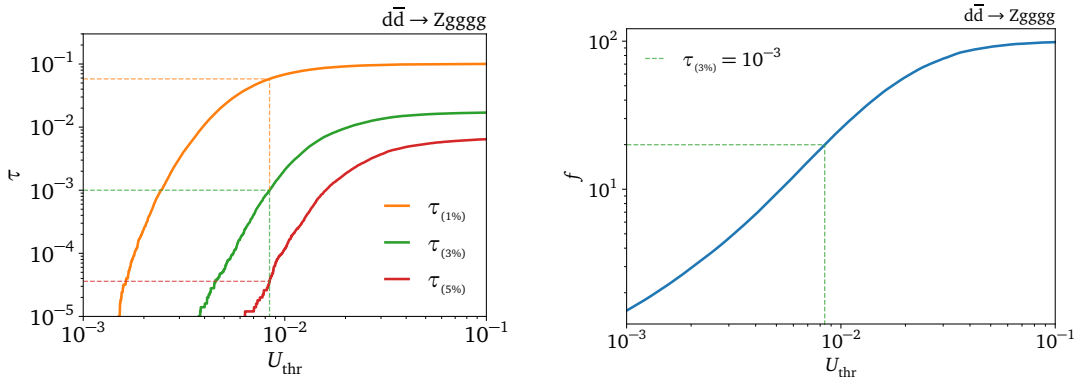

Figure 14: Tolerances (left) and Speed-up factor (right) for the ensemble [FG1^2, FG1] factorization model for the process dd̄ → Zgggg. The plots are obtained using different uncertainty thresholds for the first and second surrogate, such that: $U_1 = U_{\text{thr}}$ and $U_2 = 1.4 \cdot U_{\text{thr}}$

6 Conclusion and outlook

In this paper, we have compared various surrogates to speed-up the evaluation of the amplitude (i.e. the matrix-element squared summed/averaged over spin and colour), focusing on the case of LO event generation. For Z+jets processes, we have compared four types of ansatz: starting from a physics agnostic network (dubbed NoFa), which is a simple fully connected neural network, while the three others used a simplified Catani-Seymour approximation to inject additional physics, simplify the function to learn, and ultimately make the prediction more accurate.

Our study demonstrated that, for the non-optimal process we investigated, using physics-based approximations (i.e. the Catani-Seymour approximation) led to improved predictions compared to the NoFa network architecture, which directly predicted the logarithm of the amplitude. In contrast, applying the Catani–Seymour approximation twice recursively reduced accuracy to a level comparable to that achieved when the most singular pair was not selected for the splitting kernel. By training our network with a heteroscedastic loss, the network not

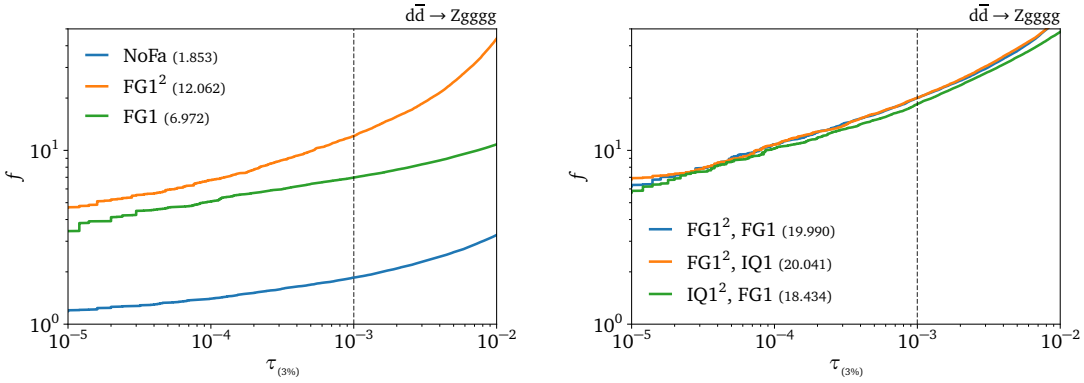

Figure 15: Speed-up factor comparison between different individual factorization models (left), and ensemble factorization models (right) for the process dd̄ → Zgggg. The values in parentheses are the speed-up factors corresponding for $\tau_{(3\%)} = 10^{-3}$.

only predicted the amplitude but also provided an estimate of its confidence in the prediction. We observed that this confidence was generally pessimistic: the network tended to overestimate the error, predicting low precision even when achieving high precision. Notably, no bias was observed, allowing us to use this estimate conservatively.

First, we leveraged both the error estimates and the correlation matrix to combine predictions from different networks. Although the predictions were highly correlated, each network learned different regions of phase space, yielding a tangible gain in precision. Second, we used surrogates to replace amplitude evaluations when the (estimated) surrogate accuracy was sufficiently precise. This mixed approach allowed full control over the induced error for any observable computed from LO samples. Considering other sources of uncertainty (PDF and scale variations), we deemed an additional error at the percent level acceptable for all observables and proposed a simple, conservative algorithm to ensure surrogate usage does not compromise predictions and remains below other sources of uncertainty.

Our results revealed a complex trade-off between approximation speed and accuracy. The optimal strategy was neither the fastest (NoFa) nor the most accurate (FG1), but rather the double-factorization network. This network, approximately 100 times faster than a full QFT computation, offers a substantial speed-up that compensates for its sub-optimal accuracy. For the conservative benchmark we selected (which correspond to an conservative additional statistical error of one percent), we achieved a speed-up factor of 6.5 for the least accurate process ($dg \to Zdggg$) and an impressive factor 20 for the process ($d\bar{d} \to Zgggg$).

Consequently, this work highlights the importance of predicting not only the amplitude but also its associated error. This capability is essential for improving surrogate accuracy and for practical applications, as it enables control over the additional numerical uncertainty introduced by surrogates. Finally, we emphasize that the goal of surrogates should not be to achieve maximal accuracy (which cannot exceed the exact computation), but to maximize speed-up for a given target accuracy, determined by the problem at hand. Thus, network speed is as critical as the prediction precision and the prediction of the error itself.

This work serves as a proof of concept. Future efforts will focus on refactoring the approach to make these surrogates accessible to users, enabling them to assess the required sample size to achieve effective gains, even when accounting for training time. Additionally, it will be important to simplify network training, accelerate evaluation, and improve the reliability of error estimation to achieve even higher speed-up factors.

Acknowledgements

We would like to thank Luigi Favaro, Tilman Plehn, Jesse Thaler, Víctor Bresó-Pla and Alessandra Fanfani for the useful discussions. Our research is funded by FRS-FNRS (Belgian National Scientific Research Fund) IISN projects 4.4503.16 (MaxLHC). This article/publication is based upon work from COST Action CA24146 and CA22130, supported by COST (European Cooperation in Science and Technology). Computational resources have been provided by the supercomputing facilities of the Université catholique de Louvain (CISM/UCL) and the Consortium des Équipements de Calcul Intensif en Fédération Wallonie Bruxelles (CÉCI) funded by the Fond de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under convention 2.5020.11 and by the Walloon Region.

A Catani-Seymour factorization formulas

Catani-Seymour factorization formulas for final state radiation

Given a process with n particles in the final state, in the singular limit of a final state radiation $p_i\,p_j \to 0$, we can express the amplitudes as [47]

$$\langle|\mathcal{M}_n|^2\rangle = \sum_{k\neq i,j} \mathcal{D}_{ij,k} + \mathcal{O}(p_i\,p_j) \tag{28}$$

where $\mathcal{O}(p_i\,p_j)$ represents non-singular terms in $p_i\,p_j \to 0$, k is a final state spectator and $\mathcal{D}_{ij,k}$ represents a dipole contribution of the form

$$\mathcal{D}_{ij,k} = -\frac{1}{2p_i\,p_j}\left\langle \mathcal{M}_{n-1}\left|\frac{\mathcal{T}_k\cdot\mathcal{T}_{ij}}{\mathcal{T}_{ij}^2}\mathcal{V}_{ij,k}\right|\mathcal{M}_{n-1}\right\rangle \tag{29}$$

where the $\mathcal{M}_{n-1}$ is the tree-level amplitude defined on the set of reduced momenta $\{(p_a,p_b\to p_1,\ldots,\tilde{p}_{ij},\tilde{p}_k,\ldots,p_{n-1})\}$. $\mathcal{T}_K$ and $\mathcal{T}_{ij}$ are the colour charges of the emitter and spectator, and $\mathcal{V}_{ij,k}$ are splitting matrices in the helicity space of the emitter. The splitting matrices, and the reduced momenta, depend on the radiation variables

$$y_{ij,k} = \frac{p_ip_j}{p_jp_i+p_jp_k+p_ip_k} \qquad \text{and} \qquad z_{ij,k} = \frac{p_ip_k}{p_ip_k+p_jp_k} \tag{30}$$

The momenta of the emitter $\tilde{p}_{ij}$ and the spectator $\tilde{p}_k$ of the reduced process are then given by

$$\tilde{p}_{ij} = p_i + p_j - \frac{y_{ij,k}}{1-y_{ij,k}}p_k \qquad \text{and} \qquad \tilde{p}_k = \frac{1}{1-y_{ij,k}}\,p_k\,. \tag{31}$$

For a quark (or anti-quark) splitting into a quark (or anti-quark) and a gluon, $q_f \to q_f + g_f$, we have

$$\langle s|\mathcal{V}_{q_ig_j,k})|s'\rangle = 8\pi\alpha C_F\left[\frac{2}{1-z_{ij,k}\,(1-y_{ij,k})} - (1+z_{ij,k}) - \epsilon(1-z_{ij,k})\right]\delta^{ss'} \tag{32}$$

where s,s' are the spin indices of the fermion $\tilde{q}_{ij}$, α is the strong coupling constant and $C_F = \frac{4}{3}$, and ϵ is a dimensional regularization parameter such that $d = 4-2\epsilon$. For a gluon splitting into a pair of gluons, $g_f \to g_f + g_f$, we have

$$\begin{aligned}\langle\mu|\mathcal{V}_{g_ig_j,k}|\nu\rangle = 16\pi\alpha C_A\Bigg[&-g^{\mu\nu}\left(\frac{1}{1-z_{ij,k}\,(1-y_{ij,k})} + \frac{1}{1-(1-z_{ij,k})\,(1-y_{ij,k})} - 2\right)\\ &+(1-\epsilon)\frac{1}{p_i\,p_j}\left(z_{ij,k}p_i^\mu - (1-z_{ij,k})p_j^\mu\right)\left(z_{ij,k}p_i^\nu - (1-z_{ij,k})p_j^\nu\right)\Bigg],\end{aligned} \tag{33}$$

where $C_A = 3$. For a gluon splitting into a pair of quark and anti-quark, $g_f \to q_f + \bar{q}_f$, we have

$$\langle\mu|\mathcal{V}_{q_i\bar{q}_j,k}|\nu\rangle = 8\pi\alpha T_R\left[-g^{\mu\nu} - \frac{2}{p_i\,p_j}\left(z_{ij,k}p_i^\mu - (1-z_{ij,k})p_j^\mu\right)\left(z_{ij,k}p_i^\nu - (1-z_{ij,k})p_j^\nu\right)\right], \tag{34}$$

where $T_R = \frac{1}{2}$.

Catani-Seymour factorization formulae for initial state radiation

In the singular limit of an initial state radiation $p_a\, p_i \to 0$, we can express the amplitude as [47]

$$\langle|\mathcal{M}_n|^2\rangle = \sum_{k\neq i} \mathcal{D}_{ai,k} + \mathcal{O}(p_a\, p_i) \tag{35}$$

where $\mathcal{D}_{ai,k}$ represents a dipole contribution of the form

$$\mathcal{D}_{ij,k} = -\frac{1}{2p_a\, p_i}\, \frac{1}{x_{ai,k}}\, \langle \mathcal{M}_{n-1}| \frac{\mathcal{T}_k \cdot \mathcal{T}_{ai}}{\mathcal{T}_{ai}^2} \mathcal{V}_{ai,k} |\mathcal{M}_{n-1}\rangle \tag{36}$$

In this case, the momenta are redefined as:

$$(p_a, p_b \to p_1, \ldots, p_i, p_k, \ldots, p_{n+1}) \Longrightarrow \{(\tilde{p}_a, p_b \to p_1, \ldots, \tilde{p}_k, \ldots, p_n)\}_{(a,i,k)} \tag{37}$$

The associated radiation variables are given by

$$x_{ai,k} = \frac{p_k p_a + p_i p_a - p_i p_k}{(p_k + p_i)\, p_a} \qquad \text{and} \qquad u_{ai,k} = \frac{p_i p_a}{p_i p_a + p_k p_a}\,. \tag{38}$$

The momenta of the emitter $\tilde{p}_a$ and the spectator $\tilde{p}_k$ in the reduced process are given by

$$\tilde{p}_a = x_{ai,k}\, p_a \qquad \text{and} \qquad \tilde{p}_k = p_k + p_i - (1 - x_{ai,k})\, p_a\,. \tag{39}$$

For a quark (or anti-quark) splitting into a quark (or anti-quark) and a gluon, $q_i \to q_i + g_f$, we have

$$\langle s|\mathcal{V}_{q_a g_i,k}|s'\rangle = 8\pi\alpha C_F \left[\frac{2}{1 - x_{ai,k} + u_{ai,k}} - (1 + x_{ai,k}) - \epsilon(1 - x_{ai,k})\right] \delta^{ss'} \tag{40}$$

For a gluon splitting into a pair of gluons, $g_i \to g_i + g_f$, the splitting function is

$$\begin{aligned} \langle \mu|\mathcal{V}_{g_a g_i,k}|\nu\rangle = 16\pi\alpha C_A \Bigg[& -g^{\mu\nu}\left(\frac{1}{1 - x_{ai,k} + u_{ai,k}} - 1 + x_{ai,k}\,(1 - x_{ai,k})\right) \\ & + (1-\epsilon)\frac{1 - x_{ai,k}}{x_{ai,k}}\, \frac{u_{ai,k}(1 - u_{ai,k})}{p_i\, p_j} \left(\frac{p_i^\mu}{u_{ai,k}} - \frac{p_k^\mu}{1 - u_{ai,k}}\right)\left(\frac{p_i^\nu}{u_{ai,k}} - \frac{p_k^\nu}{1 - u_{ai,k}}\right)\Bigg] \end{aligned} \tag{41}$$

Factorization ansatz for final-state radiation

In our factorization ansatz, we approximate the Catani–Seymour dipole by replacing the spin-correlation tensor with its spin-contracted scalar analogue

$$-\frac{1}{2p_i\, p_j} \langle \mu|\mathcal{V}_{ij,k}|\nu\rangle \longrightarrow F_{ij,k} \tag{42}$$

thereby removing the explicit helicity structure. This eliminates off-diagonal spin correlations and yields a purely scalar splitting kernel. The colour operator $\mathcal{T}_k \mathcal{T}_{ij}$ is replaced by an effective scalar colour factor $C_{ij,k}$, corresponding to the leading colour approximation.

After summing the colour–spin–correlated Born matrix element over spin and colour indices, the dipole contribution reduces to the simplified form

$$D_{ij,k} = \langle|\mathcal{M}_{n-1}|^2\rangle \cdot C_{ij,k} \cdot F_{ij,k} \tag{43}$$

In our network we implement only a single dipole factorization and do not sum over all possible spectators, avoiding the sum in Eq.(28). Furthermore, since the ansatz is ultimately multiplied by a neural-network–predicted correction factor, we neglect the explicit colour factor $C_{ij,k}$ allowing the network to absorb this dependence. In this way, our approximation for the full squared amplitude becomes:

$$\langle|\mathcal{M}_n|^2\rangle \longrightarrow \langle|\mathcal{M}_{n-1}|^2\rangle \cdot F_{ij,k} \tag{44}$$

The splitting function $F^r_{ij,k}$ depends on the specific type of radiation. For a quark (or anti-quark) splitting into a quark (or anti-quark) and a gluon, $q_f \to q_f g_f$, we have

$$F^{q_f\to q_f g_f}_{ij,k} = \frac{4\pi\alpha C_F}{p_i p_j}\left[\frac{2}{1-z_{ij,k}(1-y_{ij,k})} - 1 + z_{ij,k}\right] \tag{45}$$

where α is the strong coupling constant and $C_F = \frac{4}{3}$. For a gluon splitting into a pair of gluons, $g_f \to g_f g_f$, we have:

$$F^{g_f\to g_f g_f}_{ij,k} = \frac{8\pi\alpha C_A}{p_i p_j}\left[\frac{1}{1-z_{ij,k}(1-y_{ij,k})} + \frac{1}{1-(1-z_{ij,k})(1-y_{ij,k})} - 2\right], \tag{46}$$

where $C_A = 3$. For a gluon splitting into a pair of quark and anti-quark, $g_f \to q_f \bar{q}_f$, we have

$$F^{g_f\to q_f \bar{q}_f}_{ij,k} = \frac{4\pi\alpha T_R}{p_i p_j}, \tag{47}$$

where $T_R = \frac{1}{2}$.

Factorization ansatz for initial-state radiation

In the case of initial state radiation, the momenta are redefined as:

$$(p_a, p_b \to p_1, \ldots, p_i, p_k, \ldots, p_{n+1}) \Longrightarrow \{(\tilde{p}_a, p_b \to p_1, \ldots, \tilde{p}_k, \ldots, p_n)\}_{(a,k)} \tag{48}$$

The associated radiation variables are given by

$$x_{ai,k} = \frac{p_k p_a + p_i p_a - p_i p_k}{(p_k + p_i) p_a} \qquad \text{and} \qquad u_{ai,k} = \frac{p_i p_a}{p_i p_a + p_k p_a}. \tag{49}$$

The momenta of the emitter $\tilde{p}_a$ and the spectator $\tilde{p}_k$ in the reduced process are given by

$$\tilde{p}_a = x_{ai,k}\, p_a \qquad \text{and} \qquad \tilde{p}_k = p_k + p_i - (1 - x_{ai,k})\, p_a\,. \tag{50}$$

For a quark (or anti-quark) splitting into a quark (or anti-quark) and a gluon, $q_i \to q_i g_f$, we have

$$F^{q_i\to q_i g_f}_{ai,k} = \frac{8\pi\alpha C_F}{x_{ai,k} p_i p_i}\left[\frac{2}{1-x_{ai,k}+u_{ai,k}} - 1 - x_{ai,k}\right] \tag{51}$$

For a gluon splitting into a pair of gluons, $q_i \to q_i g_f$, the splitting function is

$$F^{q_i\to q_i g_f}_{ai,k} = \frac{8\pi\alpha C_A}{x_{ai,k} p_i p_i}\left[\frac{1}{1-x_{ai,k}+u_{ai,k}} - 1 + x_{ai,k}(1-x_{ai,k})\right] \tag{52}$$

References

[1] J. M. Campbell *et al.*, *Event Generators for High-Energy Physics Experiments*, in *Snowmass 2021*. 3, 2022. arXiv:2203.11110 [hep-ph].

[2] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands, *An introduction to PYTHIA 8.2*, Comput. Phys. Commun. **191** (2015) 159, arXiv:1410.3012 [hep-ph].

[3] Sherpa Collaboration, *Event Generation with Sherpa 2.2*, SciPost Phys. **7** (2019) 3, 034, arXiv:1905.09127 [hep-ph].

[4] J. Bellm *et al.*, *Herwig 7.1 Release Note*, arXiv:1705.06919 [hep-ph].

[5] 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*, JHEP **07** (2014) 079, arXiv:1405.0301 [hep-ph].

[6] O. Bolinder, R. Frederix, and M. Sjodahl, *All-gluon amplitudes with off-shell recursion in multiplet bases*, arXiv:2507.22636 [hep-ph].

[7] HEP Software Foundation, J. Albrecht *et al.*, *A Roadmap for HEP Software and Computing R&D for the 2020s*, Comput. Softw. Big Sci. **3** (2019) 1, 7, arXiv:1712.06982 [physics.comp-ph].

[8] J. Alwall, P. Demin, S. de Visscher, R. Frederix, M. Herquet, F. Maltoni, T. Plehn, D. L. Rainwater, and T. Stelzer, *MadGraph/MadEvent v4: The New Web Generation*, JHEP **09** (2007) 028, arXiv:0706.2334 [hep-ph].

[9] K. Hagiwara, J. Kanzaki, N. Okamura, D. Rainwater, and T. Stelzer, *Calculation of HELAS amplitudes for QCD processes using graphics processing unit (GPU)*, Eur. Phys. J. C **70** (2010) 513, arXiv:0909.5257 [hep-ph].

[10] K. Hagiwara, J. Kanzaki, N. Okamura, D. Rainwater, and T. Stelzer, *Fast calculation of HELAS amplitudes using graphics processing unit (GPU)*, Eur. Phys. J. C **66** (2010) 477, arXiv:0908.4403 [physics.comp-ph].

[11] K. Hagiwara, J. Kanzaki, Q. Li, N. Okamura, and T. Stelzer, *Fast computation of MadGraph amplitudes on graphics processing unit (GPU)*, Eur. Phys. J. C **73** (2013) 2608, arXiv:1305.0708 [physics.comp-ph].

[12] S. Hageboeck, T. Childers, W. Hopkins, O. Mattelaer, N. Nichols, S. Roiser, J. Teig, A. Valassi, C. Vuosalo, and Z. Wettersten, *Madgraph5_aMC@NLO on GPUs and vector CPUs Experience with the first alpha release*, EPJ Web Conf. **295** (2024) 11013, arXiv:2312.02898 [physics.comp-ph].

[13] Z. Wettersten, O. Mattelaer, S. Roiser, A. Valassi, and M. Zaro, *Hardware acceleration for next-to-leading order event generation within MadGraph5_aMC@NLO*, EPJ Web Conf. **337** (2025) 01230, arXiv:2503.07439 [hep-ph].

[14] A. Valassi, T. Childers, S. Hageböck, D. Massaro, O. Mattelaer, N. Nichols, F. Optolowicz, S. Roiser, J. Teig, and Z. Wettersten, *Madgraph on GPUs and vector CPUs: towards production (The 5-year journey to the first LO release CUDACPP v1.00.00)*, in *27th International Conference on Computing in High Energy and Nuclear Physics*. 3, 2025. arXiv:2503.21935 [physics.comp-ph].

[15] S. Hageböck, D. Massaro, O. Mattelaer, S. Roiser, A. Valassi, and Z. Wettersten, *Data-parallel leading-order event generation in MadGraph5_aMC@NLO*, arXiv:2507.21039 [hep-ph].

[16] E. Bothmann, T. Childers, W. Giele, S. Höche, J. Isaacson, and M. Knobbe, *A Portable Parton-Level Event Generator for the High-Luminosity LHC*, SciPost Phys. **17** (2024) 081, arXiv:2311.06198 [hep-ph].

[17] T. Plehn, A. Butter, B. Dillon, and C. Krause, *Modern Machine Learning for LHC Physicists*, arXiv:2211.01421 [hep-ph].

[18] S. Badger *et al.*, *Machine learning and LHC event generation*, SciPost Phys. **14** (3, 2023) 079, arXiv:2203.07460 [hep-ph].

[19] J. Bendavid, *Efficient Monte Carlo Integration Using Boosted Decision Trees and Generative Deep Neural Networks*, arXiv:1707.00028 [hep-ph].

[20] M. D. Klimek and M. Perelstein, *Neural Network-Based Approach to Phase Space Integration*, SciPost Phys. **9** (10, 2020) 053, arXiv:1810.11509 [hep-ph].

[21] I.-K. Chen, M. D. Klimek, and M. Perelstein, *Improved Neural Network Monte Carlo Simulation*, SciPost Phys. **10** (9, 2021) 023, arXiv:2009.07819 [hep-ph].

[22] C. Gao, J. Isaacson, and C. Krause, *i-flow: High-dimensional Integration and Sampling with Normalizing Flows*, Mach. Learn. Sci. Tech. **1** (1, 2020) 045023, arXiv:2001.05486 [physics.comp-ph].

[23] E. Bothmann, T. Janßen, M. Knobbe, T. Schmale, and S. Schumann, *Exploring phase space with Neural Importance Sampling*, SciPost Phys. **8** (1, 2020) 069, arXiv:2001.05478 [hep-ph].

[24] C. Gao, S. Höche, J. Isaacson, C. Krause, and H. Schulz, *Event Generation with Normalizing Flows*, Phys. Rev. D **101** (2020) 7, 076002, arXiv:2001.10028 [hep-ph].

[25] R. Winterhalder, V. Magerya, E. Villa, S. P. Jones, M. Kerner, A. Butter, G. Heinrich, and T. Plehn, *Targeting multi-loop integrals with neural networks*, SciPost Phys. **12** (12, 2022) 129, arXiv:2112.09145 [hep-ph].

[26] E. Bothmann, T. Janßen, M. Knobbe, B. Schmitzer, and F. Sinz, *Efficient many-jet event generation with Flow Matching*, arXiv:2506.18987 [hep-ph].

[27] T. Janßen, R. Poncelet, and S. Schumann, *Sampling NNLO QCD phase space with normalizing flows*, arXiv:2505.13608 [hep-ph].

[28] T. Heimel, R. Winterhalder, A. Butter, J. Isaacson, C. Krause, F. Maltoni, O. Mattelaer, and T. Plehn, *MadNIS – Neural Multi-Channel Importance Sampling*, arXiv:2212.06172 [hep-ph].

[29] T. Heimel, N. Huetsch, F. Maltoni, O. Mattelaer, T. Plehn, and R. Winterhalder, *The MadNIS reloaded*, SciPost Phys. **17** (2024) 1, 023, arXiv:2311.01548 [hep-ph].

[30] T. Heimel, O. Mattelaer, T. Plehn, and R. Winterhalder, *Differentiable MadNIS-Lite*, SciPost Phys. **18** (2025) 1, 017, arXiv:2408.01486 [hep-ph].

[31] F. Bishara and M. Montull, *(Machine) Learning Amplitudes for Faster Event Generation*, arXiv:1912.11055 [hep-ph].

[32] S. Badger and J. Bullock, *Using neural networks for efficient evaluation of high multiplicity scattering amplitudes*, JHEP **06** (2020) 114, arXiv:2002.07516 [hep-ph].

[33] J. Aylett-Bullock, S. Badger, and R. Moodie, *Optimising simulations for diphoton production at hadron colliders using amplitude neural networks*, JHEP **08** (6, 2021) 066, arXiv:2106.09474 [hep-ph].

[34] D. Maître and H. Truong, *A factorisation-aware Matrix element emulator*, JHEP **11** (7, 2021) 066, arXiv:2107.06625 [hep-ph].

[35] D. Maître and H. Truong, *One-loop matrix element emulation with factorisation awareness*, arXiv:2302.04005 [hep-ph].

[36] V. Bresó, G. Heinrich, V. Magerya, and A. Olsson, *Interpolating amplitudes*, arXiv:2412.09534 [hep-ph].

[37] K. Danziger, T. Janßen, S. Schumann, and F. Siegert, *Accelerating Monte Carlo event generation – rejection sampling using neural network event-weight estimates*, SciPost Phys. **12** (9, 2022) 164, arXiv:2109.11964 [hep-ph].

[38] T. Janßen, D. Maître, S. Schumann, F. Siegert, and H. Truong, *Unweighting multijet event generation using factorisation-aware neural networks*, SciPost Phys. **15** (2023) 3, 107, arXiv:2301.13562 [hep-ph].

[39] T. Herrmann, T. Janßen, M. Schenker, S. Schumann, and F. Siegert, *Accelerating multijet-merged event generation with neural network matrix element surrogates*, arXiv:2506.06203 [hep-ph].

[40] J. M. Villadamigo, R. Frederix, T. Plehn, T. Vitos, and R. Winterhalder, *FASTColor – Full-color Amplitude Surrogate Toolkit for QCD*, arXiv:2509.07068 [hep-ph].

[41] J. Spinner, V. Bresó, P. de Haan, T. Plehn, J. Thaler, and J. Brehmer, *Lorentz-Equivariant Geometric Algebra Transformers for High-Energy Physics*, in *38th conference on Neural Information Processing Systems*. 10, 2024. arXiv:2405.14806 [physics.data-an].

[42] J. Brehmer, V. Bresó, P. de Haan, T. Plehn, H. Qu, J. Spinner, and J. Thaler, *A Lorentz-equivariant transformer for all of the LHC*, SciPost Phys. **19** (2025) 4, 108, arXiv:2411.00446 [hep-ph].

[43] L. Favaro, G. Gerhartz, F. A. Hamprecht, P. Lippmann, S. Pitz, T. Plehn, H. Qu, and J. Spinner, *Lorentz-Equivariance without Limitations*, arXiv:2508.14898 [hep-ph].

[44] S. Badger, A. Butter, M. Luchmann, S. Pitz, and T. Plehn, *Loop Amplitudes from Precision Networks*, SciPost Phys. Core **6** (6, 2023) 034, arXiv:2206.14831 [hep-ph].

[45] H. Bahl, N. Elmer, L. Favaro, M. Haußmann, T. Plehn, and R. Winterhalder, *Accurate Surrogate Amplitudes with Calibrated Uncertainties*, SciPost Phys. Core **8** (2025) 073, arXiv:2412.12069 [hep-ph].

[46] H. Bahl, N. Elmer, T. Plehn, and R. Winterhalder, *Amplitude Uncertainties Everywhere All at Once*, arXiv:2509.00155 [hep-ph].

[47] S. Catani and M. H. Seymour, *A General algorithm for calculating jet cross-sections in NLO QCD*, Nucl. Phys. B **485** (1997) 291, arXiv:hep-ph/9605323. [Erratum: Nucl.Phys.B 510, 503–504 (1998)].

[48] S. Catani, S. Dittmaier, M. H. Seymour, and Z. Trocsanyi, *The Dipole formalism for next-to-leading order QCD calculations with massive partons*, Nucl. Phys. B **627** (2002) 189, arXiv:hep-ph/0201036.

[49] D. MacKay, *Probable Networks and Plausible Predictions – A Review of Practical Bayesian Methods for Supervised Neural Networks*, Comp. in Neural Systems **6** (1995) 4679.

[50] R. M. Neal, *Bayesian learning for neural networks*. PhD thesis, Toronto, 1995. ftp://www.cs.toronto.edu/dist/radford/thesis.pdf.

[51] Y. Gal, *Uncertainty in Deep Learning*. PhD thesis, Cambridge, 2016. http://mlg.eng.cam.ac.uk/yarin/thesis/thesis.pdf.

[52] S. Bollweg, M. Haußmann, G. Kasieczka, M. Luchmann, T. Plehn, and J. Thompson, *Deep-Learning Jets with Uncertainties and More*, SciPost Phys. **8** (2020) 1, 006, arXiv:1904.10004 [hep-ph].

[53] G. Kasieczka, M. Luchmann, F. Otterpohl, and T. Plehn, *Per-Object Systematics using Deep-Learned Calibration*, SciPost Phys. **9** (3, 2020) 089, arXiv:2003.11099 [hep-ph].

[54] ATLAS Collaboration, *Precision calibration of calorimeter signals in the ATLAS experiment using an uncertainty-aware neural network*, arXiv:2412.04370 [hep-ex].

[55] J. Yi and M. A. Bessa, *Cooperative bayesian and variance networks disentangle aleatoric and epistemic uncertainties*, arXiv:2505.02743 [cs.LG].

[56] L. Lyons, D. Gibaut, and P. Clifford, *How to Combine Correlated Estimates of a Single Physical Quantity*, Nucl. Instrum. Meth. A **270** (1988) 110.

[57] A. Valassi, *Combining correlated measurements of several different physical quantities*, Nucl. Instrum. Meth. A **500** (2003) 391.

[58] G. P. Lepage, *A New Algorithm for Adaptive Multidimensional Integration*, J. Comput. Phys. **27** (1978) 192.

[59] G. P. Lepage, *VEGAS: AN ADAPTIVE MULTIDIMENSIONAL INTEGRATION PROGRAM*, .

[60] G. P. Lepage, *Adaptive multidimensional integration: VEGAS enhanced*, J. Comput. Phys. **439** (2021) 110386, arXiv:2009.05112 [physics.comp-ph].

[61] O. Mattelaer and K. Ostrolenk, *Speeding up MadGraph5_aMC@NLO*, Eur. Phys. J. C **81** (2021) 5, 435, arXiv:2102.00773 [hep-ph].

[62] F. Maltoni and T. Stelzer, *MadEvent: Automatic event generation with MadGraph*, JHEP **02** (2003) 027, arXiv:hep-ph/0208156.

[63] A. van Hameren, *PARNI for importance sampling and density estimation*, Acta Phys. Polon. B **40** (2009) 259, arXiv:0710.2448 [hep-ph].

[64] R. Kleiss, W. J. Stirling, and S. D. Ellis, *A New Monte Carlo Treatment of Multiparticle Phase Space at High-energies*, Comput. Phys. Commun. **40** (1986) 359.

[65] R. Kleiss and R. Pittau, *Weight optimization in multichannel Monte Carlo*, Comput. Phys. Commun. **83** (1994) 141, arXiv:hep-ph/9405257.

[66] E. Bothmann, T. Childers, W. Giele, F. Herren, S. Hoeche, J. Isaacson, M. Knobbe, and R. Wang, *Efficient phase-space generation for hadron collider event simulation*, SciPost Phys. **15** (2023) 4, 169, arXiv:2302.10449 [hep-ph].

[67] L. Kish, *Survey Sampling*. John Wiley & Sons, 1965.

[68] G. M. Amdahl, *Validity of the single processor approach to achieving large scale computing capabilities*, in *Proceedings of the Spring Joint Computer Conference (AFIPS '67)*. ACM, Atlantic City, NJ, USA, April 18–20, 1967.