



# MadFlow: automating Monte Carlo simulation on GPU for particle physics processes

Stefano Carrazza<sup>1,2,3,a</sup>, Juan Cruz-Martinez<sup>1</sup>, Marco Rossi<sup>1,2</sup>, Marco Zaro<sup>1</sup>

<sup>1</sup> TIF Lab, Dipartimento di Fisica, Università degli Studi di Milano and INFN Sezione di Milano, Milan, Italy

<sup>2</sup> CERN Theoretical Physics Department and OpenLab, 1211 Geneva 23, Switzerland

<sup>3</sup> Quantum Research Centre, Technology Innovation Institute, Abu Dhabi, United Arab Emirates

Received: 18 June 2021 / Accepted: 10 July 2021  
© The Author(s) 2021

**Abstract** We present MadFlow, a first general multi-purpose framework for Monte Carlo (MC) event simulation of particle physics processes designed to take full advantage of hardware accelerators, in particular, graphics processing units (GPUs). The automation process of generating all the required components for MC simulation of a generic physics process and its deployment on hardware accelerator is still a big challenge nowadays. In order to solve this challenge, we design a workflow and code library which provides to the user the possibility to simulate custom processes through the MadGraph5\_aMC@NLO framework and a plugin for the generation and exporting of specialized code in a GPU-like format. The exported code includes analytic expressions for matrix elements and phase space. The simulation is performed using the VegasFlow and PDFFlow libraries which deploy automatically the full simulation on systems with different hardware acceleration capabilities, such as multi-threading CPU, single-GPU and multi-GPU setups. The package also provides an asynchronous unweighted events procedure to store simulation results. Crucially, although only Leading Order is automatized, the library provides all ingredients necessary to build full complex Monte Carlo simulators in a modern, extensible and maintainable way. We show simulation results at leading-order for multiple processes on different hardware configurations.

## 1 Introduction

The popularity of hardware accelerators, such as graphics processing units (GPU), has quickly increased in the last years thanks to the exceptional performance benefits and efficiency achieved in scientific and industrial applications. Furthermore, new code frameworks based on hardware accelerators

have been designed in order to simplify the implementation of algorithms and models in particular in the context of Artificial Intelligence applications.

If we consider the research domain of High Energy Physics, we can observe several examples of applications that could benefit from the conversion or systematic implementation of existing algorithms and code libraries on GPU. Some examples have already been successfully published, such as deep learning applications in experimental physics [1], where astonishing performance improvements were obtained thanks to the employment of GPUs.

Despite the HEP community interest in providing computational tools for experimental setups, we still observe a growing trend towards the increase of computational time required to solve complex problems [2,3] in particular for Monte Carlo (MC) simulation of particle physics processes. Moreover, this growing trend is further increased by the current state of the art implementations of MC simulation libraries, which still today rely almost exclusively, on CPU architecture [4–7]. This is despite the fact that the parallel nature of MC simulations makes them the perfect target for hardware accelerators. Attempts at porting MC simulation libraries to GPUs have shown quite promising results, but they have been limited in scope [8–12].

From a practical perspective, in order to write a competitive GPU-capable full parton-level MC by any measure with existing tools, there are at least five required ingredients: (i) an integrator, able to parallelize over the number of events; (ii) a GPU-capable parton distribution function (PDF) interpolation tool; (iii) an efficient phase space generator, which should generate valid phase space points on GPU, apply any fiducial cuts; (iv) finally evaluate the matrix element squared for the target processes; (v) an efficient asynchronous output storage system for observables, such as histograms and Les Houches event files.

<sup>a</sup> e-mail: [stefano.carrazza@cern.ch](mailto:stefano.carrazza@cern.ch) (corresponding author)

In recent previous works, we have developed open-source tools that provide the ground basis for the implementation of an automatic Monte Carlo simulation framework for HEP addressing some of the aforementioned issues: VegasFlow [13, 14] and PDFFlow [15, 16]. The first package, VegasFlow, is a new software for fast evaluation of high dimensional integrals based on Monte Carlo integration techniques designed for platforms with hardware accelerators. This project allows developers to delegate all complicated aspects of hardware or platform implementation to the library, reducing development time due to maintainability and debugging. On the other hand, PDFFlow is a library which provides fast evaluation of parton distribution functions (PDFs) designed for platforms with hardware accelerators following the design idea inspired from VegasFlow. The availability of both packages completes respectively points (i) and (ii) above.

The goal of this work is to address and propose an integrated technical solution for all points above, following the original idea presented in [17]. We call `MadFlow` [18] the open-source software library which implements this automatic pipeline for GPU deployment of Monte Carlo simulation. It combines the matrix elements expressions generated by the `MadGraph5_aMC@NLO` (`MG5_aMC`) [5, 6] framework with the VegasFlow and PDFFlow efficient simulation tool for hardware accelerators. `MadFlow` by design, opens the possibility to study and benchmark multiple approaches to Monte Carlo integration based on distributed hardware, and, in future, new algorithms based on deep learning techniques.

This work is structured as follows. In Sect. 2 we describe the technical implementation of `MadFlow`. In Sect. 3 we compare and benchmark results. Finally, in Sect. 4 we present our conclusion and future development roadmap.

## 2 Implementation

### 2.1 The `MadFlow` concept

`MadFlow` is an open-source python package [18], which provides the user a simple tool for parton-level Monte Carlo simulation on hardware accelerators. The original concept of `MadFlow` is to keep usability and maintainability as simple as possible thanks to the new technologies and frameworks currently available today.

From the user's point of view, the effort and time required to start using `MadFlow` and write a fully-working Leading Order Monte Carlo simulation is limited to the installation of the package and its dependencies. The library provides a quick-start script which takes care of the generation of all required modules and code dependencies for running the simulation on the user's hardware, including multi-threading

CPU, single GPU and multi-GPU setups for both NVIDIA and AMD products.

On the other hand, from the developer perspective, the `MadFlow` code is based on primitives actively maintained by the community and its design is modular: all components presented in the next paragraphs can be modified and extended with minor effort, thanks to an uniform standardization of the frameworks and documentation.

### 2.2 The `MadFlow` code design

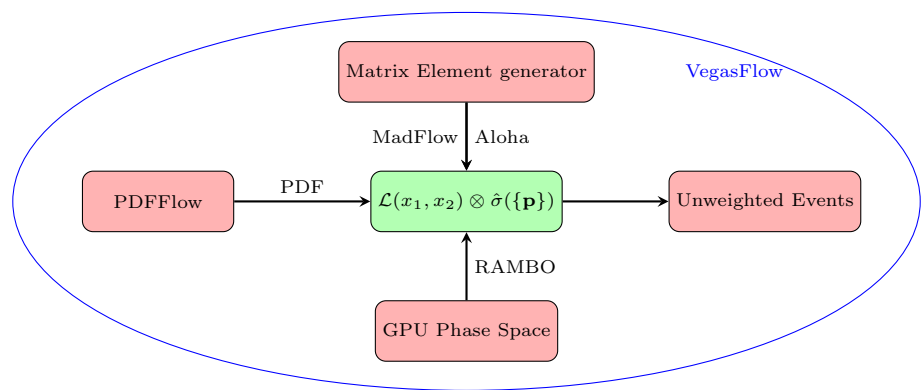
Nowadays many research groups rely on very extensive and complex code bases, thus learning how to use an equally complicated framework might require time and expertise which may not be feasible for everyone. For example, consider the investment in time required for training of new doctoral students or researchers. Therefore, in order to accelerate the adoption of hardware accelerators within the hep-ph community we design the `MadFlow` Monte Carlo implementation on GPU with maintainability and developer-friendliness as a major target feature.

We consider the `MG5_aMC` framework as the entry point of our procedure. `MG5_aMC` is a meta-code written in Python, that generates automatically the code in a low-level language to be employed for the simulation of arbitrary scattering processes at colliders, in the Standard Model or beyond. `MG5_aMC` relies on general procedures and methods, without being tailored to a specific process, class of processes, or physics model. Besides the generation of tree-level matrix elements, `MG5_aMC` gives also the possibility to the user to include Next-to-Leading order corrections, both due to strong and electroweak interactions (including matching to parton-shower for the former). However, in this paper we will limit ourselves to the case of tree-level matrix elements.

The workflow of `MG5_aMC` is the following: a model, written in the Universal Feynrules Output (UFO) format [19], is loaded, which contains all the informations on the underlying theory, including the Feynman rules. Starting from the model, Feynman diagrams are generated, and the corresponding expressions for the matrix elements are written in process-specific files. The various parts of the Feynman diagrams (external wavefunctions, vertices, propagators, etc.) are evaluated via the ALOHA routines [20] (with the introduction of `MadGraph5` [21] ALOHA supersedes the HELAS routines [22]).

It is then natural to consider `MG5_aMC`, in particular ALOHA, as a backend of matrix elements for GPU for a future general purpose parton-level GPU MC generator. We should note there is an independent effort dedicated to porting `MG5_aMC` to GPU [23]. The `MadFlow` project differs from "Madgraph 4 GPU" for two main reasons: the interest in providing a full MC simulator based on modern software which can automatically be deployed in different hardware

**Fig. 1** Schematic workflow for the implementation of MadFlow. The software automates the generation of a custom process using the standard MG5\_aMC framework API and exports the relevant code in a specific format for VegasFlow and PDFFlow integration



configurations and the need of a technical solution which simplifies maintainability and does not require specialized GPU knowledge from the developer and user point of view. However, we must note CUDA-based libraries are compatible with TensorFlow and thus it is technically possible to use “Madgraph 4 GPU” matrix elements within MadFlow, maximizing the advantages both codes provide.

In Fig. 1 we show the modules involved in the current implementation of MadFlow. The process is driven by a MadFlow script which generates a custom process using the MG5\_aMC standard framework API and exports the relevant code for the analytic matrix elements and phase space expressions in python, using the syntax defined by VegasFlow. In terms of code implementation this step has required the development of a MG5\_aMC plugin, which consists of an exporter module to write the matrix element and the ALOHA routines in Python, fulfilling the requirements imposed by VegasFlow and PDFFlow using TensorFlow [24] primitives. The main difficulty consists in converting sequential functions into vectorized functions. During the MadFlow development we have performed several numerical and performance benchmarks in order to avoid potential bugs.

After the conversion step into the specified format is performed, the exported python code is incorporated into a VegasFlow device-agnostic integration algorithm which executes the event generation pipeline from the generation of random numbers, computation of the phase space kinematics, matrix element evaluation and histogram accumulation.

### 2.3 The evaluation of matrix elements routines

In MadFlow, the matrix elements evaluations follows the original MG5\_aMC implementation: a `Matrix` class is produced by the Python exporter plugin module. Its `smatrix` method links together the needed Feynman rules to compute the requested matrix element: it loops over initial and final state particle helicities and aggregates their contribution to the final squared amplitude.

The matrix element vectorization requires to replace the ALOHA waveforms and vertices routines abiding by the Ten-

sorFlow ControlFlow rules. Although this process is straightforward for vertices Feynman rules, being mostly comprised by elementary algebraic operations, the implementation of particle waveforms functions is subject to several conditional control statements that make the task harder as GPUs suffer considerably from branching.

### 2.4 Phase-space generation

The integration phase-space is generated using a vectorized implementation of the RAMBO algorithm [25] which makes it suitable for hardware accelerators.

RAMBO maps the random variables of the integrator library to a flat phase-space with which the `smatrix` method of the matrix element can be evaluated. While this means MadFlow can produce results for any number of particles, the generated phase space doesn’t take into account the topology of the process. As a result, for a great number of final-state particles the number of events required to get a reasonably precise result is much larger than what would be required with other Monte Carlos.

More complex and efficient phase-space generators will be developed for future releases of MadFlow.

### 2.5 Unweighted events exporter

The Madflow event generator is equipped with a Les Houches Event (LHE) writer module that provides a class to store events in the LHE 3.0 file format [26]. The LHE writer class operates asynchronously in a separated thread from the VegasFlow integration, thus ensuring that the integrator computational performance is not harmed by IO limitations. The module works by collecting all the unweighed events generated by the integrator and applying an unweighting procedure employing the module provided by MG5\_aMC. The final result is a (compressed) LHE file, with unweighted events.

We note that in this implementation, however, the unweighting efficiency is rather low (around 5%) because of the non-optimised phase space which relies on RAMBO.

## 2.6 Scope

The goal of `MadFlow` is to provide the foundation for future high precision Monte Carlo simulation (of higher orders or otherwise) so they can efficiently take advantage of hardware developments.

In its current form, `MadFlow` provides the necessary tools for the computation of Leading Order (LO) calculations fully automatically for any number of particles.<sup>1</sup> Higher order calculations can be implemented by building upon the provided LO template. Parameters and model definitions are provided by `MG5_aMC`-compatible parameters card and so the same options and models can be used or defined.

An important caveat must be considered when trying to integrate more complex processes. The provided phase-space is flat and does not take into account the topology of the diagrams (see Sect. 2.4) and therefore becomes inefficient with large multiplicities. As a result a very large number of events might be necessary to reduce the Monte Carlo error, cancelling the benefits of running on a GPU. Therefore, for processes with many particles in the final state or when using the tree level amplitudes for Next-to-Leading Order (NLO) calculations writing an specialized phase-space is recommended [27]. For some recent developments on this matter, see also [28].

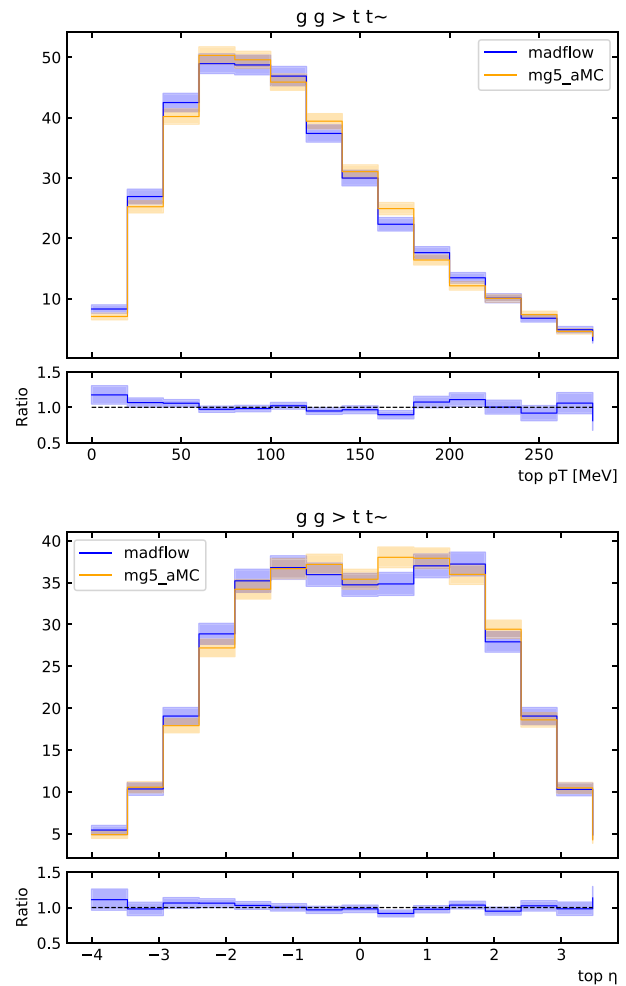
In summary, while due to some of its limitations `MadFlow` cannot provide yet efficient calculations, it can quickstart the development of efficient fixed-order Monte Carlo simulators in hardware accelerators.

## 3 Results

In this section we present accuracy and performance results of `MadFlow` on consumer and professional grade GPU and CPU devices. We focus on Leading Order calculation for hadronic processes at  $\sqrt{s} = 13$  TeV. For this exercise we select 5 processes with growing number of diagrams and channels in order to emulate the behaviour in terms of complexity of a Next-to-Leading Order and Next-to-Next to Leading Order computations. In particular, we consider:  $gg \rightarrow t\bar{t}$  (3 diagrams),  $pp \rightarrow t\bar{t}$  (7 diagrams),  $pp \rightarrow t\bar{t}g$  (36 diagrams),  $pp \rightarrow t\bar{t}gg$  (267 diagrams) and  $pp \rightarrow t\bar{t}ggg$  (2604 diagrams).

Note that all results presented in this section have been computed using the matrix elements generated by `MG5_aMC` in double precision without any further optimization, thus further releases of `MadFlow` will address systematically memory and performance optimization in order to achieve even better results.

<sup>1</sup> The code is still in beta testing and some corner cases may fail due to lack of tests.

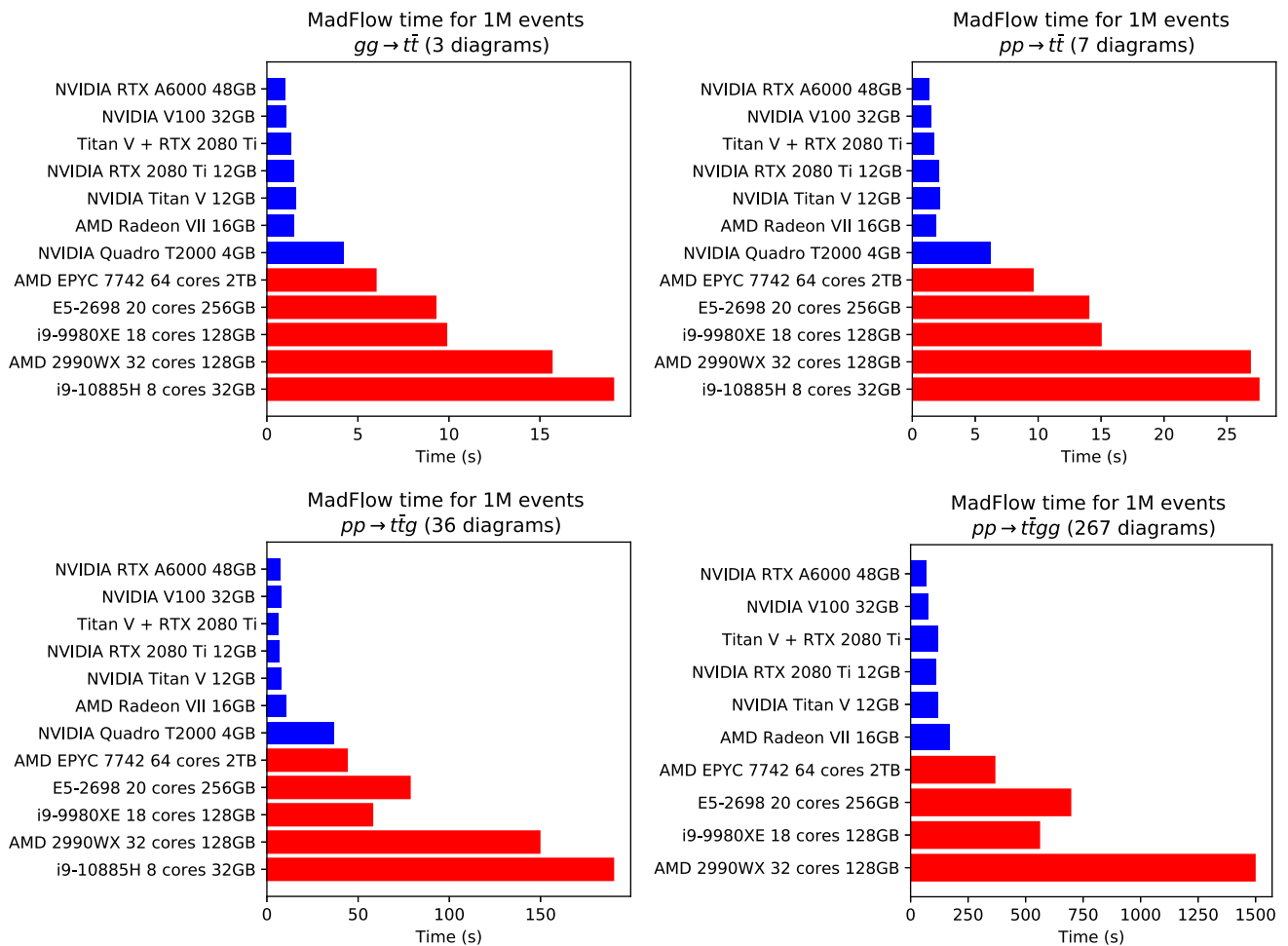


**Fig. 2** Leading Order cross section differential on  $p_{t,\text{top}}$  (first row) and  $\eta_{\text{top}}$  (second row) for  $gg \rightarrow t\bar{t}$  at  $\sqrt{s} = 13$  TeV. We compare predictions between `MadFlow` (blue) and `MG5_aMC` (orange). We observe that in both cases the distributions are in statistical agreement

All results presented in this section are obtained with `madflow` 0.1, `vegasflow` 1.2.1, `pdfflow` 1.2.1, `MG5_aMC` 3.1.0, `tensorflow` 2.5.0 for NVIDIA/Intel/EPYC systems with CUDA 11.3 drivers, and `tensorflow-rocm` 2.4.1 with ROCm 4.2.0 drivers on Radeon/AMD systems.

### 3.1 Accuracy

In Fig. 2 we show an example of Leading Order cross section differential on  $p_{t,\text{top}}$  and  $\eta_{\text{top}}$  for  $gg \rightarrow t\bar{t}$  at  $\sqrt{s} = 13$  TeV for predictions obtained with the original `MG5_aMC` integration procedure and the `MadFlow` approach based on `VegasFlow` and `PDFFlow`. In the first row we show the differential distribution in  $p_{t,\text{top}}$  using the absolute scale in fb/GeV and the respective ratio between both MC predictions, while in the second row we show the  $\eta_{\text{top}}$  distribution, confirming a



**Fig. 3** Timings obtained with MadFlow to evaluate events at Leading Order for  $gg \rightarrow t\bar{t}$  (top left),  $pp \rightarrow t\bar{t}$  (top right),  $pp \rightarrow t\bar{t}g$  (bottom left) and  $pp \rightarrow t\bar{t}gg$  (bottom right). We show results for consumer

and professional grade GPUs (blue bars) and CPUs (red bars). For each device we quote the available RAM memory. We observe a systematic performance advantage for GPU devices

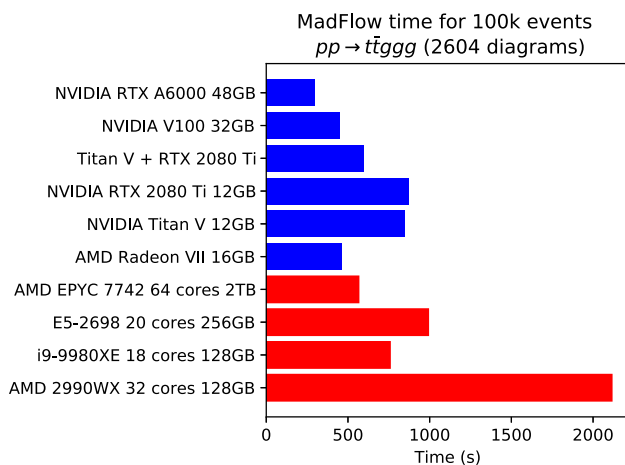
good level of agreement between both implementations for the same level of target accuracy between 2-5% for each bin.

The results presented here are computed independently from each framework in order to minimize communication bottlenecks between CPU-GPU. The plots are constructed from unweighted events stored using the LHE approach described in Sect. 2.5.

### 3.2 Performance

In terms of performance, in particular evaluation time, in Fig. 3 we compare the total amount of time required by MadFlow for the computation of 1M events for the processes described above:  $gg \rightarrow t\bar{t}$  (3 diagrams),  $pp \rightarrow t\bar{t}$  (7 diagrams),  $pp \rightarrow t\bar{t}g$  (36 diagrams),  $pp \rightarrow t\bar{t}gg$  (267 diagrams). For all simulations, we apply a  $p_t > 30$  GeV cut for all out-going particles. We performed the simulation on multiple Intel and AMD CPU configurations (red bars), together

with NVIDIA and AMD GPUs (blue bars) ranging from consumer to professional grade hardware. Blue bars show the greatest performance of MadFlow when running on GPU devices. We observe that NVIDIA GPUs with the Ampere architecture, such as the RTX A6000, out-performs the previous Tesla generation. We have observed that the performance of the AMD Radeon VII is comparable to most professional grade GPUs presented in the plot. The red bars show the timings for the same code evaluated on CPU using all available cores. We confirm that GPU timings are quite competitive when compared to CPU performance, however some top-level chips such as the AMD Epyc 7742, can get similar performance results when compared to general consumer level GPUs, such as the Quadro T2000. Note that in order to obtain good performance and going into production mode, the MadFlow user should adjust the maximum number of events per device, in order to occupy the maximum amount of memory available. We conclude that the MadFlow imple-



**Fig. 4** Same as Fig. 3 for  $pp \rightarrow t\bar{t}gg$  at Leading Order. We confirm that a large number of diagrams can be deployed on GPU and obtain relevant performance improvements when compared to CPU results

**Table 1** Comparison of event computation time for MadFlow and MG5\_aMC, using an Intel i9-9980XE with 18 cores and 128 GB of RAM for CPU simulation and the NVIDIA Titan V 12 GB for GPU simulation

Process	MadFlow CPU ( $\mu$ s)	MadFlow GPU ( $\mu$ s)	MG5_aMC ( $\mu$ s)
$gg \rightarrow t\bar{t}$	9.86	1.56	20.21
$pp \rightarrow t\bar{t}$	14.99	2.20	45.74
$pp \rightarrow t\bar{t}g$	57.84	7.54	93.23
$pp \rightarrow t\bar{t}gg$	559.67	121.05	793.92

mentation confirms a great performance improvement when running on GPU hardware, providing an interesting trade-off in terms of price cost and generated events.

In Fig. 4 we present a preliminary example of simulation timings for 100k events using MadFlow as described above for  $pp \rightarrow t\bar{t}gg$  with 2604 diagrams. The code generated for this example follows the same procedure adopted for processes shown in Fig. 3. We can remarkably confirm that MadFlow results on GPU are competitive when compared CPU results even for a such large number of diagrams (and thus required GPU memory), taking into account that no custom code optimization has been included. It is certainly true that the memory footprint and the overall performances of the code can (and should) be improved, e.g. by considering the Leading-Color approximation of the matrix element and/or possibly by performing a Monte-Carlo over color and helicity configurations, we believe that these results confirm that GPU computation has a strong potential in HEP simulations at higher orders.

### 3.3 Comparing to MG5\_aMC

Finally, in Table 1 we measure and compare the required time per event for the processes discussed above using MadFlow and MG5\_aMC simulations on a Intel i9-9980XE CPU with 18 cores and 128GB of RAM and a NVIDIA Titan V 12GB GPU. As expected, we confirm that MadFlow on GPU increases dramatically the evaluated number of events per second.

Finally, as expected, the performance gain for GPUs when compared to CPU decreases with the number of diagrams included in a given process thanks to the amount of memory required to hold the computation workload. Such limitation could be partially improved by using GPU models with larger memory, e.g. the new NVIDIA A100 with 80GB, by compressing and optimizing the kernel codes before execution [12,23], and by using multi-GPU configurations where portions of diagrams are distributed across devices.

## 4 Outlook

In conclusion in this work we present MadFlow, a new approach for the generalization of Monte Carlo simulation on hardware accelerators. In particular, the MadFlow design provides a fast and maintainable code which can quickly port complex analytic expressions into hardware specific languages without complex operations involving several computer languages, tools and compilers. Furthermore, we confirm the algorithm effectiveness when running simulation on hardware accelerators.

The MadFlow code is open-source and public available on GitHub<sup>2</sup> [18]. The repository contains links to documentation for installation, hardware setup, examples and development.

As an outlook, we plan to continue the development of MadFlow as an open-source library. Foreseen major improvements include: to replace the RAMBO phase-space with more efficient solutions based on the process topology; to investigate the possibility to accelerate integration using machine learning techniques; finally, to set the stage for the implementation of all required changes to accommodate Next-to-Leading order computations.

**Acknowledgements** We thank the authors of “Madgraph 4 GPU”, Olivier Mattelaer, Stefan Roiser and Andrea Valassi for useful discussions. M.Z. wishes to thank Fabio Maltoni for discussions and suggestions. S.C. and J.C.M are supported by the European Research Council under the European Union’s Horizon 2020 research and innovation Programme, grant agreement number 740006. M.Z. is supported by Programma per Giovani Ricercatori “Rita Levi Montalcini” granted by the

<sup>2</sup> <https://github.com/N3PDF/madflow>

Italian Ministero dell'Università e della Ricerca (MUR). This project is supported by the UNIMI Linea 2A 2021 program.

**Data Availability Statement** This manuscript has no associated data or the data will not be deposited. [Authors' comment: The paper presents a tool for fast simulation on hardware accelerators.]

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. Funded by SCOAP<sup>3</sup>.

## References

1. K. Albertsson et al., J. Phys. Conf. Ser. **1085**, 022008 (2018). [arXiv:1807.02876](https://arxiv.org/abs/1807.02876)
2. J. Niehues, D.M. Walker, Phys. Lett. B **788**, 243 (2019). [arXiv:1807.02529](https://arxiv.org/abs/1807.02529)
3. S. Hoche et al., Working group report: computing for perturbative QCD, in *Community Summer Study 2013: Snowmass on the Mississippi* (2013). [arXiv:1309.3598](https://arxiv.org/abs/1309.3598)
4. T. Gleisberg, S. Hoeche, F. Krauss, M. Schonherr, S. Schumann, F. Siegert, J. Winter, JHEP **02**, 007 (2009). [arXiv:0811.4622](https://arxiv.org/abs/0811.4622)
5. J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H.S. Shao, T. Stelzer, P. Torrielli, M. Zaro, JHEP **07**, 079 (2014). [arXiv:1405.0301](https://arxiv.org/abs/1405.0301)
6. R. Frederix, S. Frixione, V. Hirschi, D. Pagani, H.S. Shao, M. Zaro, JHEP **07**, 185 (2018). [arXiv:1804.10017](https://arxiv.org/abs/1804.10017)
7. J. Campbell, T. Neumann, JHEP **12**, 034 (2019). [arXiv:1909.09117](https://arxiv.org/abs/1909.09117)
8. K. Hagiwara, J. Kanzaki, N. Okamura, D. Rainwater, T. Stelzer, Eur. Phys. J. C **66**, 477 (2010). [arXiv:0908.4403](https://arxiv.org/abs/0908.4403)
9. K. Hagiwara, J. Kanzaki, N. Okamura, D. Rainwater, T. Stelzer, Eur. Phys. J. C **70**, 513 (2010). [arXiv:0909.5257](https://arxiv.org/abs/0909.5257)
10. K. Hagiwara, J. Kanzaki, Q. Li, N. Okamura, T. Stelzer, Eur. Phys. J. C **73**, 2608 (2013). [arXiv:1305.0708](https://arxiv.org/abs/1305.0708)
11. G. Grasseau, F. Beaudette, C. MartinPerez, A. Zabi, A. Chiron, T. Strebler, G. Hautreux, EPJ Web Conf. **214**, 06028 (2019)
12. E. Bothmann, W. Giele, S. Hoeche, J. Isaacson, M. Knobbe (2021). [arXiv:2106.06507](https://arxiv.org/abs/2106.06507)
13. S. Carrazza, J.M. Cruz-Martinez, Comput. Phys. Commun. **254**, 107376 (2020). [arXiv:2002.12921](https://arxiv.org/abs/2002.12921)
14. S. Carrazza, J.M. Cruz-Martinez, VegasFlow: accelerating Monte Carlo simulation across platforms, in *40th International Conference on High Energy Physics* (2020). [arXiv:2010.09341](https://arxiv.org/abs/2010.09341)
15. S. Carrazza, J.M. Cruz-Martinez, M. Rossi (2020). [arXiv:2009.06635](https://arxiv.org/abs/2009.06635)
16. M. Rossi, S. Carrazza, J.M. Cruz-Martinez, PDFFlow: hardware accelerating parton density access, in *40th International Conference on High Energy Physics* (2020). [arXiv:2012.08221](https://arxiv.org/abs/2012.08221)
17. S. Carrazza, J. Cruz-Martinez, M. Rossi, M. Zaro, Towards the automation of Monte Carlo simulation on GPU for particle physics processes, in *25th International Conference on Computing in High-Energy and Nuclear Physics* (2021). [arXiv:2105.10529](https://arxiv.org/abs/2105.10529)
18. S. Carrazza, J.M. Cruz-Martinez, M. Rossi, M. Zaro, *N3PDF/madflow* (2021). <https://doi.org/10.5281/zenodo.4954376>. <https://github.com/N3PDF/madflow>
19. C. Degrande, C. Duhr, B. Fuks, D. Grellscheid, O. Mattelaer, T. Reiter, Comput. Phys. Commun. **183**, 1201 (2012). [arXiv:1108.2040](https://arxiv.org/abs/1108.2040)
20. P. de Aquino, W. Link, F. Maltoni, O. Mattelaer, T. Stelzer, Comput. Phys. Commun. **183**, 2254 (2012). [arXiv:1108.2041](https://arxiv.org/abs/1108.2041)
21. J. Alwall, M. Herquet, F. Maltoni, O. Mattelaer, T. Stelzer, JHEP **06**, 128 (2011). [arXiv:1106.0522](https://arxiv.org/abs/1106.0522)
22. H. Murayama, I. Watanabe, K. Hagiwara (1992)
23. *Madgraph 4 gpu*. <https://github.com/madgraph5/madgraph4gpu>
24. M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin et al., *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems* (2015). Software available from tensorflow.org. <http://tensorflow.org/>
25. R. Kleiss, W.J. Stirling, S.D. Ellis, Comput. Phys. Commun. **40**, 359 (1986)
26. J. Butterworth, G. Dissertori, S. Dittmaier, D. de Florian, N. Glover, K. Hamilton, J. Huston, M. Kado, A. Korytov, F. Krauss et al., *Les Houches 2013: Physics at TeV Colliders: Standard Model Working Group Report* (2014). [arXiv:1405.1067](https://arxiv.org/abs/1405.1067)
27. R. Kleiss, R. Pittau, Comput. Phys. Commun. **83**, 141 (1994). [arXiv:hep-ph/9405257](https://arxiv.org/abs/hep-ph/9405257)
28. O. Mattelaer, K. Ostrolenk, Eur. Phys. J. C **81**, 435 (2021). [arXiv:2102.00773](https://arxiv.org/abs/2102.00773)