AtlFast3: Fast Simulation in ATLAS for LHC Run 3 and Beyond

AtlFast3: Fast Simulation in ATLAS for LHC Run 3 and Beyond

Federico Andrea Guillaume Corchia[1,2], *Marilena* Bandieramonte[3], *Joshua Falco* Beirer[4], *John Derek* Chapman[5], *Michael* Dührssen-Debling[4], *Florian* Ernst[4,6], *Michele* Faucci Giannelli[7], *Tong* Qiu[8], *Jana* Schaarschmidt[9], *Firdaus* Soberi[8], and *Rui* Zhang[10] on behalf of the ATLAS Computing Activity

[1]Department of Physics and Astronomy, University of Bologna, Bologna, Italy
[2]INFN Bologna, Bologna, Italy
[3]Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, United States
[4]CERN
[5]Cavendish Laboratory, University of Cambridge, Cambridge, United Kingdom
[6]Institut für Theoretische Physik, Universität Heidelberg, Heidelberg, Germany
[7]Chalmers University of Technology, Gothenburg, Sweden
[8]SUPA - School of Physics and Astronomy, University of Edinburgh, Edinburgh, United Kingdom
[9]Department of Physics, University of Washington, Seattle, United States
[10]Department of Physics, Nanjing University, Nanjing, China

Abstract. As we are approaching the high-luminosity era of the LHC, the computational requirements of the ATLAS experiment are expected to increase significantly in the coming years. Notably, simulation of Monte Carlo (MC) events is immensely computationally demanding, and their limited availability is one of the major sources of systematic uncertainties in many physics analyses. The main bottleneck in detector simulation is the detailed simulation of electromagnetic and hadronic showers in the ATLAS calorimeter system using Geant4.
To increase MC statistics and to leverage the available CPU resources for LHC Run 3, the ATLAS Collaboration has recently put into production a refined and significantly improved version of its state-of-the-art fast simulation tool AtlFast3. AtlFast3 uses classical parametric and machine learning-based approaches such as Generative Adversarial Networks (GANs) for fast simulation of LHC events in the ATLAS detector.
This work presents the newly improved version of AtlFast3 that is currently in production for simulation of Run 3 samples. In addition, ideas and plans for the future of fast simulation in ATLAS are also discussed.

1 Introduction

Simulation of Monte Carlo (MC) events in the detector is a major computing challenge at LHC experiments, taking about 40% of the total load on the computing resources [1] of the ATLAS experiment [2]. A large part is taken by the detailed simulation of electromagnetic

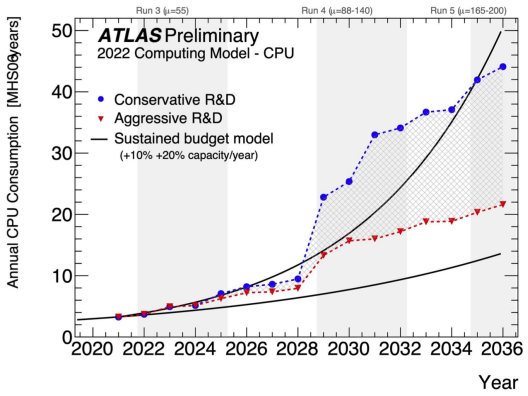

Figure 1. Projected evolution of compute usage, under a "conservative" and "aggressive" R&D scenarios. The hatched shading shows the range of resource consumption if the aggressive scenario is only partially achieved; the solid lines indicate the impact of sustained year-on-year budget increases and improvement in new hardware amounting together to a capacity increase of 10% (lower line) and 20% (upper line). The vertical shaded bands indicate the periods of data taking for ATLAS [3].

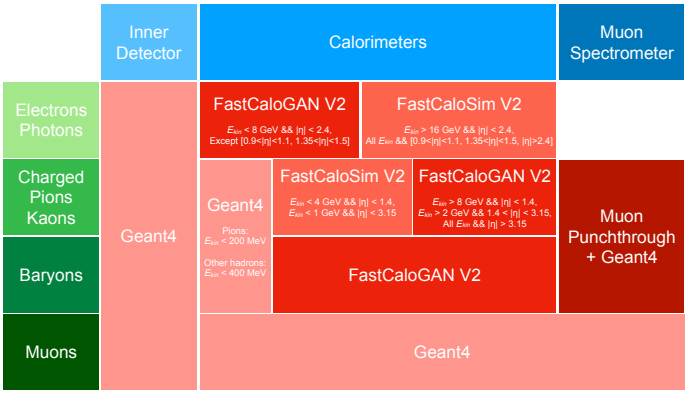

Figure 2. Configuration of the various subsystems of AtlFast3, as used for Run 3, depending on detector region, particle type and particle energy. Geant4 is still used to simulate all particles in the Inner Detector, low energy hadrons in the calorimeters and muons. Muon Punchthrough (the spray of particles into the Muon Spectrometer resulting from late interacting high-energy hadrons) is modelled with a tool based on Deep Neural Networks (DNNs) [4].

and hadronic showers in the ATLAS calorimeter system (about 80% of CPU consumption in the case of $t\bar{t}$ processes) [5]. In addition, the computational requirements of the ATLAS experiment are expected to increase significantly in the coming years, already during the current data taking run of the LHC (Run 3) and also in view of Run 4, which will be the first one of the High Luminosity LHC (HL-LHC) (see Fig. 1) [3].

These demanding and increasing requirements necessitate significant research and development and one of the solutions is the introduction of *fast simulation* tools. These are programs able to simulate the detector response faster than the standard full process simulation tool Geant4 [6–8], while keeping the loss of accuracy to a minimum.

2 AtlFast3

AtlFast3 is the fast simulation tool developed by the ATLAS Collaboration [4, 5, 9]. Introduced for Run 2, it was further improved in preparation for the current Run 3 and is now in production. The tool replaces the slow propagation and interactions of particles inside the calorimeter volume with the direct generation of energy deposits, by means of an underlying parametrisation.

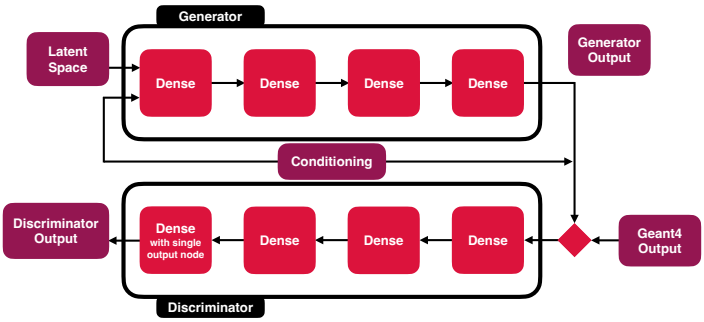

Figure 3. Architecture of FastCaloGANV2. The architecture of the network and its hyperparameters have been optimised.

The tool is composed of two subsystems, corresponding to two different approaches to fast simulation (descriptions for both are given in the coming sections):

- FastCaloSimV2, doing longitudinal and lateral parametrisation of showers;
- FastCaloGANV2, machine learning-based, employing Generative Adversarial Networks (GANs).

AtlFast3 runs fast simulation either through FastCaloSimV2 or FastCaloGANV2: the choice of the simulator was carefully tuned through extensive validation and is optimised to give the best physics performance. This configuration is shown in Fig. 2.

2.1 FastCaloSimV2

FastCaloSimV2 is a fast simulation tool that separately parametrises the longitudinal and lateral shower development [5]. During simulation, energy is directly deposited into the calorimeter cells using the parametrised responses. Parametrisation is done using Geant4 single photon, electron and charged pion samples, in 17 logarithmically spaced energy bins from 64 MeV to 4 TeV and 100 linearly spaced bins in pseudorapidity from 0 to 5 in absolute value. For the longitudinal shower development, FastCaloSimV2 considers that the amount of energy deposited in each layer by the shower particles (which depends on how deep in the calorimeter the shower was initiated) is highly correlated between layers, making it difficult to independently parametrise the response for each layer. To address this, Principal Component Analysis (PCA) is used to classify showers for each slice of energy, pseudorapidity bin and particle type. Two PCA transformations are performed: the first one classifies showers into bins, while the second one is perfomed in each bin of the first PCA to generate uncorrelated and approximately Gaussian distributions. During simulation, PCA bins are randomly selected, followed by the generation of uncorrelated random numbers, which are then mapped back to the total energy and the energy fractions deposited in each layer with the inverse transformation. The lateral shower shape is instead parametrised as two-dimensional probability density functions.

2.2 FastCaloGANV2

FastCaloGANV2 is a ML-based fast simulation system based on GANs [10]. This architecture, first introduced in [11], involves the simultaneous training of two neural networks,

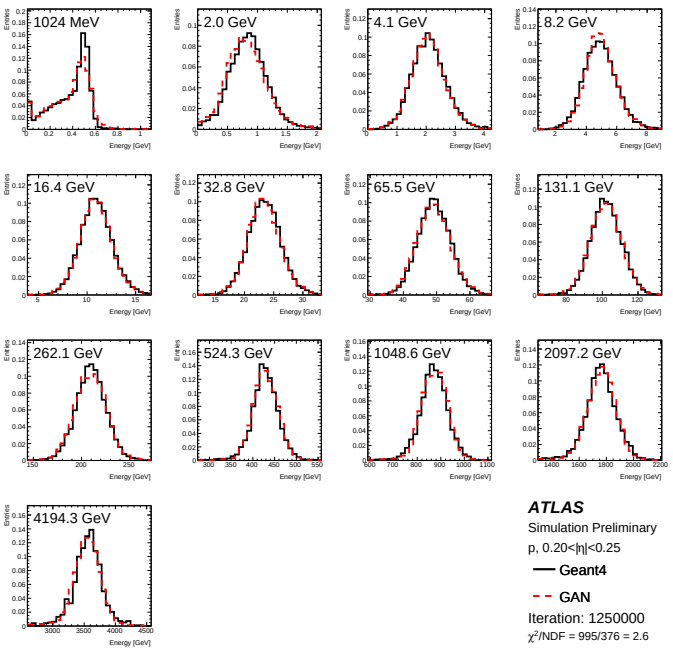

Figure 4. Sum of the energy in all voxels for single protons generated at the calorimeter surface in the pseudorapidity range between 0.2 and 0.25 in absolute value. GEANT4 is compared to the GAN trained within FASTCALOGANV2 [12].

one called *generator* and the other one called *discriminator*. The generator aims to generate samples as similar as possible to GEANT4 generated showers, while the discriminator is fed samples from both GEANT4 and the generator and aims to distinguish actual GEANT4 samples from the ones produced by the generator. Both try improving their abilities and, when a Nash equilibrium between the two is reached, the FASTCALOGANV2 generator is ready to simulate calorimeter showers and it runs much faster than GEANT4, keeping good accuracy.

In particular, FASTCALOGANV2 employs Wasserstein GANs with gradient penalty (WGAN-GP) [13], an evolution of the original GAN architecture that improves training by making it more stable and performant. The complete architecture is shown in Fig. 3. FASTCALOGANV2 was trained for electrons, photons, charged pions and protons as shower-initiating particles, in each of the 100 bins in pseudorapidity and with conditioning (the GAN being supplied additional information acting as class labels) on truth momentum. All used GEANT4 samples are produced without noise at the reconstruction level. Calorimeter hits are grouped into three-dimensional bins (*voxels*), whose granularity is optimised and finer than the one of the calorimeter cells, which improves modelling. The system is trained to reproduce voxels and energies in layers, as well as the total energy, in a single step. Fig. 4 compares simulations as done by GEANT4 and FASTCALOGANV2, with protons as shower-initiating particles; *reduced* χ^2 (the χ^2 per degree of freedom) equal to 2.6 is observed.

3 Performance

This section discusses the performance of ATLFAST3 in terms of speedup and physics results.

For what concerns speedups, it is observed that ATLFAST3 is from 3 to 15 times faster than the GEANT4 simulation of the ATLAS Run 3 detector, the lowest and highest speedup being

observed respectively for $Z \rightarrow ee$ events and for high transverse momentum dijet events (see Fig. 5). Simulation time in AtlFast3 is dominated by simulation of the Inner Detector, which, as was shown in Fig. 2, is fully handled by Geant4.

For physics results, AtlFast3 provides very accurate modelling of the leading cluster energy and, in dijet events, of the number of constituents for jet and substructure variables (see Fig. 6). For most observables used in physics analyses, AtlFast3 and Geant4 agree within a few percent and AtlFast3 physics performance is improved compared to its predecessor. AtlFast3 has also the advantage of being able to be used for almost every process, be it signal or background. The fact that AtlFast3 can cover a really large variety of analyses is particularly helpful considering that, for Run 4, more than 90-95% of analyses are expected to require the use of the fast simulation, as there will not be the CPU capacity to allow the full simulation for them [1].

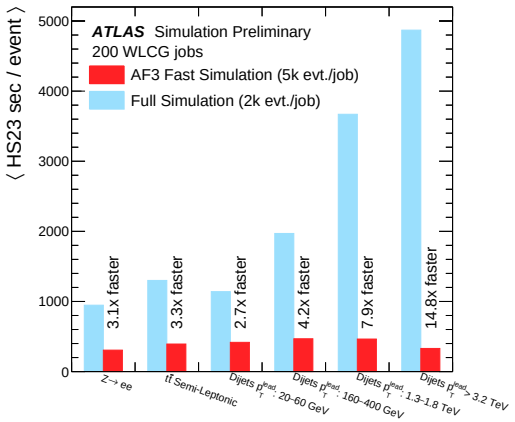

Figure 5. Mean CPU time per event simulated in AtlFast3 and full simulation measured in standardised HS23 seconds. Six different physics processes are shown; in all of them AtlFast3 is significantly faster, with the most dramatic improvement for processes with the highest energy particles [14].

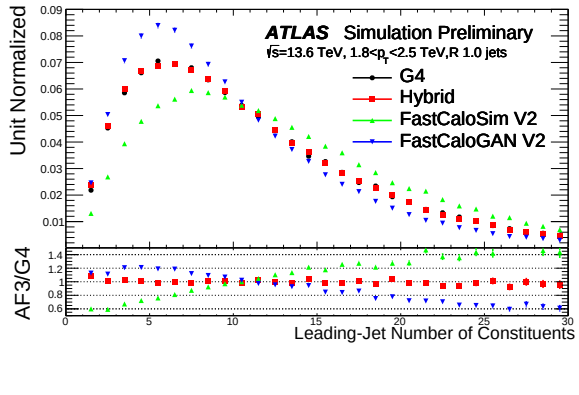

Figure 6. Number of constituents for the leading reconstructed jet in dijet events with transverse momentum between 1.8 and 2.5 TeV [12]. Results are compared for samples simulated with Geant4, FastCaloSimV2, FastCaloGANV2 and a combination of FastCaloSimV2 or FastCaloGANV2 ("Hybrid") according to the hadron energy as done in AtlFast3 for Run 2 [5].

4 Looking Ahead

After the successful insertion into the Run 3 production, further development is underway in view of the final part of Run 3 and the first part of HL-LHC, Run 4. Improved voxelisation techniques aimed at reducing bias are currently being tested, and additional ML models are under investigation - particularly diffusion models (*CaloDiT* [15, 16]) and Invertible Neural Networks (INNs [17]). These models were identified as top performers in the *CaloChallenge* study [15] and are being considered for integration into the fast simulation framework. If future developments maintain or improve their simulation quality, they may be incorporated as additional fast simulation subsystems, potentially supplementing or replacing existing components such as FastCaloSimV2 and FastCaloGANV2.

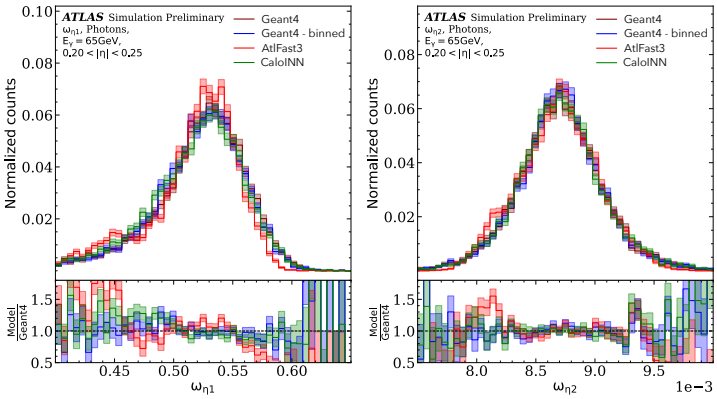

Figure 7. Reconstructed lateral shower width in the presampling layer (left) and in the second layer (right) of the barrel part of the ATLAS LAr electromagnetic calorimeter for 65 GeV photons in the pseudorapidity region between 0.20 and 0.25 in absolute value, as simulated with GEANT4 (*Geant4*), GEANT4 using artificially binned showers (*Geant4 - binned*), the nominal ATLFAST3 simulation (*AtlFast3*) and an INN trained on the binned showers (*CaloINN*). In the binned GEANT4 simulation, energy deposits are grouped into predefined spatial regions, creating an artificial discretisation [18].

The voxelisation has been re-optimised, as part of the above mentioned ongoing tests, in order to better emulate GEANT4. The new voxelisation features finer granularity at the shower centre to achieve higher spatial precision while accounting for voxel-to-voxel correlations. For the new ML models, current results show that GANs still have potential for further improvement and INNs also yield an excellent simulation performance, as they have been shown to accurately reproduce the GEANT4 distributions with the optimised voxelisation (see Fig. 7).

4.1 An Additional Tool: FASTCALOGANTAINER

Usage of FASTCALOGANV2 requires its GANs to be trained and this training requires large resources. Therefore, additional help can come from using for FASTCALOGANV2 training also other resources than the ones commonly used at CERN, like the CERN batch system LXBATCH [19] and the Worldwide LHC Computing Grid (WLCG) [20].

FASTCALOGANTAINER makes it possible for FASTCALOGANV2 training to run on unpledged resources external to CERN, without them needing to be configured as LXBATCH or the WLCG. FASTCALOGANTAINER is based on an Apptainer container [21], built starting from the Docker image of the operating system of LXBATCH, and includes the standard ATLAS software environment and the rest of the additional software required for training, for FASTCALOGANTAINER to be independent from the system where it is deployed. It requires libraries CUDA-11 [22] and CUDNN [23] for GPU usage. By using FASTCALOGANTAINER, resources are freed up and, in addition, further improvement in training speed can be obtained if training is run on more powerful resources like HPC farms.

FASTCALOGANTAINER was deployed on the resources shown in Tab. 1. Speedup for training is 3-4 times on Leonardo [24] (it can be noted the usage of NVIDIA A100 GPUs) and 2-3 times on CNAF-HPC [25], both values being with respect to LXBATCH. These results show the performance boost and the advantage usage of cutting-edge supercomputers provides.

Work is undergoing to deploy FASTCALOGANTAINER on other resources than the ones used up to now (including cloud resources), other architectures (notably ARM), and for more

Table 1. FastCaloGANTainer performance on various clusters. For pions one GAN was trained for all incident particle energy values; for photons two GANs were trained in parallel, one for incident particle energies up to 4 GeV, the other one for energies above. Photon GANs use batch normalisation, which contributes to the longer training times. Trainings ran with one GPU on the same amount of training samples and with the same hyperparameters. $\tilde{\chi}^2$ is the reduced χ^2: its value staying the same for each particle type across the different clusters guarantees invariance of physics performance. The greater speed of CNAF-HPC over LXBATCH with the same GPU model is due to training samples being on a file system mounted on the node itself on CNAF-HPC, while on LXBATCH they are hosted on a network file system, which has slower access times.

Resource	Features	$\pi^{\pm}$ Results	γ Results
LXBATCH (reference cluster)	CERN batch system. CentOS 7 (for used nodes), CVMFS, HTCondor, V100 GPUs	Runtime: 12 h $\tilde{\chi}^2 \approx 2$	Runtime: 30-31 h $\tilde{\chi}^2 \approx 5$
Leonardo	TOP500 10$^{\text{th}}$ most powerful cluster [26], at CINECA. RHEL 8.7, no CVMFS, SLURM, A100 GPUs, isolated nodes	Runtime: 3.5 h $\tilde{\chi}^2 \approx 2$	Runtime: 6.5-7.5 h $\tilde{\chi}^2 \approx 5$
CNAF-HPC	INFN-CNAF HPC cluster (close to the local WLCG Tier-1 cluster *INFN-T1* [27]). CentOS 7, no CVMFS, SLURM, V100 GPUs	Runtime: 6 h $\tilde{\chi}^2 \approx 2$	Runtime: 9-10 h $\tilde{\chi}^2 \approx 5$

types of shower-initiating particles. Code optimisation is also being studied, concerning in particular how to harness even better the computational power of multi-CPU/GPU nodes.

5 Conclusion

In this work the Run 3 configuration of AtlFast3 was presented: the state-of-the-art fast simulation tool of the ATLAS Collaboration at the LHC. AtlFast3 for Run 3 can simulate a broad range of physics processes with high precision, with improved physics performance compared to its predecessor. This comes with great improvements in computing performance, as AtlFast3 runs with a CPU speedup of a factor 3-15. This tool is therefore essential to meet the computational requirements of the future runs of the LHC, as well as the physics modelling accuracy needs. In addition, the associated system FastCaloGANTainer provides further relief to the computational resources commonly used at CERN by making it possible for FastCaloGANV2 training to also run on unpledged external resources, with a performance boost if run on cutting-edge supercomputers.

References

[1] ATLAS Collaboration, ATLAS HL-LHC Computing Conceptual Design Report. CERN-LHCC-2020-015 (2020). cds.cern.ch/record/2729668

[2] ATLAS Collaboration, The ATLAS Experiment at the CERN Large Hadron Collider. JINST **3** S08003 (2008). doi.org/10.1088/1748-0221/3/08/S08003

[3] ATLAS Collaboration, ATLAS Software and Computing HL-LHC Roadmap. CERN-LHCC-2022-005 (2022). cds.cern.ch/record/2802918

[4] ATLAS Collaboration, Software and computing for Run 3 of the ATLAS experiment at the LHC. Eur. Phys. J. C **85**, 234 (2025). doi.org/10.1140/epjc/s10052-024-13701-w

[5] ATLAS Collaboration, AtlFast3: The Next Generation of Fast Simulation in ATLAS. Comput Softw Big Sci **6**, 7 (2022). doi.org/10.1007/s41781-021-00079-7

[6] J. Allison et al., Recent Developments in Geant4. Nucl. Instrum. Meth. A **835**, 186-225 (2016). doi.org/10.1016/j.nima.2016.06.125

[7] J. Allison et al., Geant4 Developments and Applications. IEEE Trans. Nucl. Sci. **53**, 270-278 (2006). dx.doi.org/10.1109/TNS.2006.869826

[8] S. Agostinelli et al., Geant4 - A Simulation Toolkit. Nucl. Instrum. Meth. A **506**, 250-303 (2003). dx.doi.org/10.1016/S0168-9002(03)01368-8

[9] J. F. Beirer, Novel Approaches to the Fast Simulation of the ATLAS Calorimeter and Performance Studies of Track-Assisted Reclustered Jets for Searches for Resonant $X \to SH \to b\bar{b}WW^*$ Production with the ATLAS Detector (2023). dx.doi.org/10.53846/goediss-9900

[10] ATLAS Collaboration, Fast simulation of the ATLAS calorimeter system with Generative Adversarial Networks. ATL-SOFT-PUB-2020-006 (2020). cds.cern.ch/record/2746032

[11] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative Adversarial Nets. Advances in Neural Information Processing Systems **3**, 11 (2014). dx.doi.org/10.1145/3422622

[12] ATLAS Collaboration, ATLAS-SIM-2023-004. atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/SIM-2023-004

[13] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. Courville, Improved Training of Wasserstein GANs. arXiv:1704.00028 (2017). doi.org/10.48550/arXiv.1704.00028

[14] ATLAS Collaboration, ATLAS-SIM-2023-005. atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/SIM-2023-005

[15] C. Krause et al., CaloChallenge 2022: A Community Challenge for Fast Calorimeter Simulation. arXiv:2410.21611 (2024). doi.org/10.48550/arXiv.2410.21611 to be published in Rep. Prog. Phys.

[16] W. Peebles, S. Xie, Scalable Diffusion Models with Transformers. arXiv:2212.09748 (2022). doi.org/10.48550/arXiv.2212.09748

[17] F. Ernst, L. Favaro, C. Krause, T. Plehn, D. Shih, Normalizing Flows for High-Dimensional Detector Simulations. SciPost Phys. **18**, 081 (2025). doi.org/10.21468/SciPostPhys.18.3.081

[18] ATLAS Collaboration, ATLAS-SIMU-2024-10. atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/SIMU-2024-10

[19] batchdocs.web.cern.ch (accessed 10 February 2025)

[20] home.cern/science/computing/grid (accessed 20 February 2025)

[21] Singularity Developers, Singularity (2021). doi.org/10.5281/zenodo.1310023 (accessed 10 February 2025)

[22] developer.nvidia.com/cuda-zone (accessed 10 February 2025)

[23] developer.nvidia.com/cudnn (accessed 10 February 2025)

[24] leonardo-supercomputer.cineca.eu (accessed 10 February 2025)

[25] confluence.infn.it/spaces/TD/pages/40665319/6+-+The+HPC+cluster (accessed 10 February 2025)

[26] top500.org/lists/top500/2025/06 (accessed 21 July 2025)

[27] cnaf.infn.it/en/wlcg-tier-1-data-center-en (accessed 10 February 2025)