Machine Learning: Science and Technology

PAPER • **OPEN ACCESS**

Configurable calorimeter simulation for AI applications

To cite this article: Anton Charkin-Gorbulin *et al* 2023 *Mach. Learn.: Sci. Technol.* **4** 035042

View the article online for updates and enhancements.

You may also like

- Automated gadget discovery in the quantum domain
  Lea M Trenkwalder, Andrea López-Incera, Hendrik Poulsen Nautrup et al.
- Polynomial differentiation decreases the training time complexity of physics-informed neural networks and strengthens their approximation power
  Juan-Esteban Suarez Cardona and Michael Hecht
- Hierarchical Bayesian pharmacometrics analysis of Baclofen for alcohol use disorder
  Nina Baldy, Nicolas Simon, Viktor K Jirsa et al.

This content was downloaded from IP address 204.239.251.2 on 10/12/2023 at 04:22

MACHINE
LEARNING
Science and Technology

CrossMark

OPEN ACCESS

RECEIVED
14 March 2023

REVISED
31 July 2023

ACCEPTED FOR PUBLICATION
17 August 2023

PUBLISHED
5 September 2023

Original Content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence.

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

PAPER

Configurable calorimeter simulation for AI applications

Anton Charkin-Gorbulin[2], Kyle Cranmer[4,5], Francesco Armando Di Bello[1], Etienne Dreyer[3,*], Sanmay Ganguly[6,*], Eilam Gross[3], Lukas Heinrich[7], Marumi Kado[8,9], Nilotpal Kakati[3], Patrick Rieck[4,*], Lorenzo Santi[9] and Matteo Tusoni[9]

[1] INFN and University of Genova, Genova, Italy
[2] University of Luxembourg, Luxembourg, Luxembourg
[3] Weizmann Institute of Science, Rehovot, Israel
[4] New York University, New York, United States of America
[5] University of Wisconsin-Madison, Madison, Wisconsin, United States of America
[6] ICEPP, University of Tokyo, Tokyo, Japan
[7] Technical University of Munich, Munich, Germany
[8] Max Planck Institute for Physics, Munich, Germany
[9] INFN and Sapienza University of Rome, Rome, Italy
[*] Authors to whom any correspondence should be addressed.

E-mail: etienne.dreyer@weizmann.ac.il, sanmay.ganguly@cern.ch and patrick.rieck@cern.ch

Keywords: manuscript, inserted, editor, OPEN DATA, machine learning, calorimetry

Supplementary material for this article is available online

Abstract

A configurable calorimeter simulation for AI (CoCoA) applications is presented, based on the Geant4 toolkit and interfaced with the Pythia event generator. This open-source project is aimed to support the development of machine learning algorithms in high energy physics that rely on realistic particle shower descriptions, such as reconstruction, fast simulation, and low-level analysis. Specifications such as the granularity and material of its nearly hermetic geometry are user-configurable. The tool is supplemented with simple event processing including topological clustering, jet algorithms, and a nearest-neighbors graph construction. Formatting is also provided to visualise events using the Phoenix event display software.

1. Introduction

Algorithms incorporating machine learning (ML) methods are a new paradigm in reconstruction, calibration, identification, simulation and analysis of high energy physics (HEP) experimental data. In recent years, various ML architectures have been deployed to optimize low-level tasks such as clustering, reconstruction, fast simulation, pileup suppression and object identification [1, 2]. For example, ML-based fast calorimeter simulation relies on accurate target data to train a fast conditional generative model $p(D|T)$, where T denotes the true set of stable final state particles produced in the collision and D is the set of resulting detector hits. In particle reconstruction, on the other hand, the inverse process $D \rightarrow T$ is modelled by predicting a set of particles $R(D)$ to approximate T as accurately as possible. The development of such algorithms requires a realistic, highly-granular simulation of particle detector response going beyond parameterized detector models frequently used in studies of particle physics phenomenology such as DELPHES [3]. In particular due to the complexity of particle showers in calorimeters, a detailed, microscopic simulation of interactions between particles and detector material is needed in order to develop low-level ML algorithms exploiting such features.

Recent research efforts to study calorimeter shower properties using ML [4–6] made use of the GEANT4 [7] simulation toolkit for simple detector geometries. However, an open source detector simulation with a realistic cylindrical geometry and hermetic coverage is yet to be adopted by the HEP community for ML studies and beyond. Aiming to bridge this gap, the COnfigurable Calorimeter simulatiOn for Ai (CoCoA) was developed, which uses GEANT4 [7] to implement detailed shower simulation for particles in a full-coverage, highly-segmented sensitive volume comparable to that of multipurpose detectors at the LHC. CoCoA offers research teams a common, curated benchmark tool for the benefits of straightforward detector

© 2023 The Author(s). Published by IOP Publishing Ltd

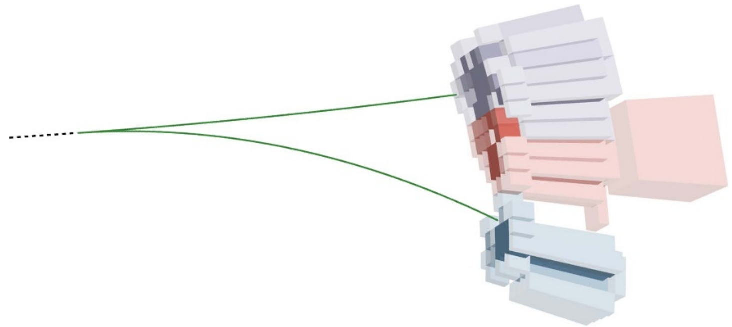


Figure 1. Visualization of a photon (dashed line) with energy 50 GeV converting to two electrons (green lines) producing three distinct clusters in the CoCoA central electromagnetic calorimeter. The cluster shown in red contains an additional cell in the first layer of the hadronic calorimeter due to a noise fluctuation. Cells are shown with an opacity proportional to energy over noise ratio divided by 4.6, the threshold for topoclustering seeds.

configuration and comparability of research results. The program source code [8] is linked together with a technical documentation on the project website[10]. The emphasis of this software package is on realistic calorimeter simulation. No realistic digitization and electronic readouts are implemented and energy loss due to these processes are neglected in this package. For the same reason, simplified tracking is included in CoCoA to model particle deflection in a magnetic field and energy depositions upstream of the calorimeter. A sophisticated open source toolkit suitable for tracking studies based on silicon hits is provided by [9].

Usability for ML-based studies is a core motivation in the design of the CoCoA code. Datasets generated by CoCoA have featured in two recent applications of ML to particle reconstruction and fast simulation [10, 11]. To this end, the main parameters of the calorimeters are largely configurable, including their material, granularity, depth and the amount of readout noise. Similarly the inclusion of material interactions in the tracking region is optional. For comparisons with benchmark reconstruction approaches, output data from CoCoA are conveniently interfaced to standard topological clustering and jet clustering algorithms. The output includes a record of energy contributions to each cell by truth particles for supervising cell-level predictions and edge lists for connecting cells and tracks in a graph to support geometric deep learning models. Finally, the default geometry has been formatted for rendering in the Phoenix event display software [12], along with a script to export event output files for visualization. An example is shown in figure 1.

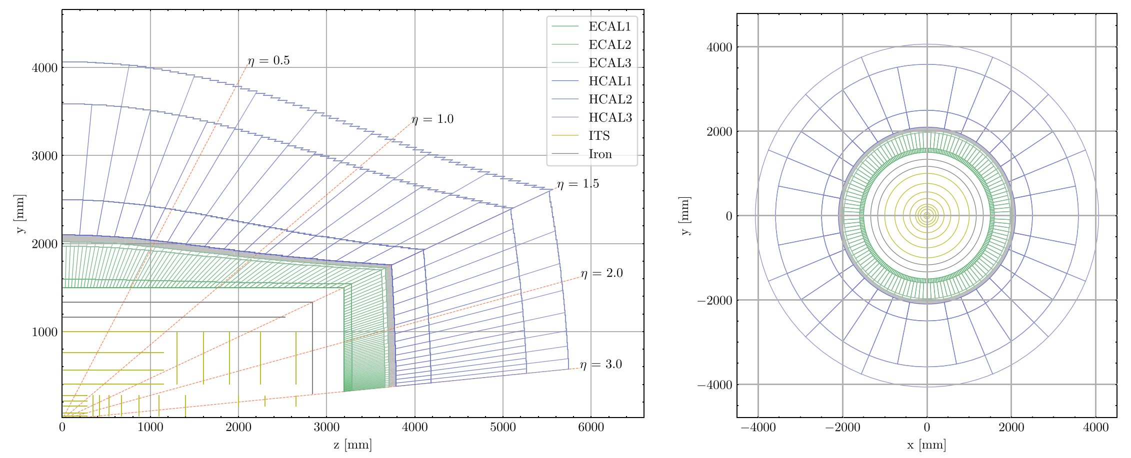
The sophisticated CoCoA calorimeter simulation and its data post-processing provides users easy access to datasets suitable to train models for current collider experiments or for more general algorithms development and benchmarking. In addition, the open-source nature of the package and its visualization support have the potential for use cases in education and science communication in HEP.

2. Detector design

The major components of CoCoA are an inner tracking system (ITS) surrounded by an electromagnetic calorimeter (ECAL) and finally a hadronic calorimeter (HCAL). These subsystems are arranged concentrically and are symmetric in azimuthal angle $-\pi < \phi \leqslant \pi$ as shown in figure 2. No muon spectrometer is considered in this design and muons are reconstructed as tracks with the ITS. The goal of this design is to accurately model the relevant outputs of a multipurpose detector at the LHC while being simplified by the exclusion of detailed components like readout electronics, cabling, and support structures. The detector design is largely configurable, with its default parameter values chosen to achieve response characteristics comparable to that of the current ATLAS detector. Following is a detailed description of each subsystem.

The ITS consists of hollow cylinders in the central detector part and disks at both of its ends, each of which are centered around the beamline. Each of these components consists of a silicon layer of 150 μm thickness in case of the disks and the five innermost cylinders and 320 μm in case of the 4 outermost cylinders. Each silicon layer is accompanied by an iron layer of 350 μm thickness in order to provide a simulation for support material. The ITS only serves the purpose of simulating the interaction of particles with matter upstream of the calorimeter. The resulting detector hits are not used for tracking purposes. The default value of the magnetic flux density present in the ITS amounts to 3.8 T. Finally, two layers of iron totaling 4.4 cm in depth are added to represent support or cryostat material in front of the calorimeter.

[10] https://cocoa-hep.readthedocs.io/en/latest/index.html.



(a) (b)

Figure 2. Positive quadrant scheme of CoCoA. We use a right-handed orthogonal coordinate system *x*-*y*-*z*, where *z*-axis is the principal axis of the detector and a constant *z* refers to a circular cross-section of the detector. (a) *yz*-projection showing the CoCoA ITS, subsequent iron layers, calorimeter system in the barrel and end-cap region, overlaid on lines marking constant pseudorapidity η. (b) *xy*-projection shows the barrel region of the same subsystems at $z = 0$.

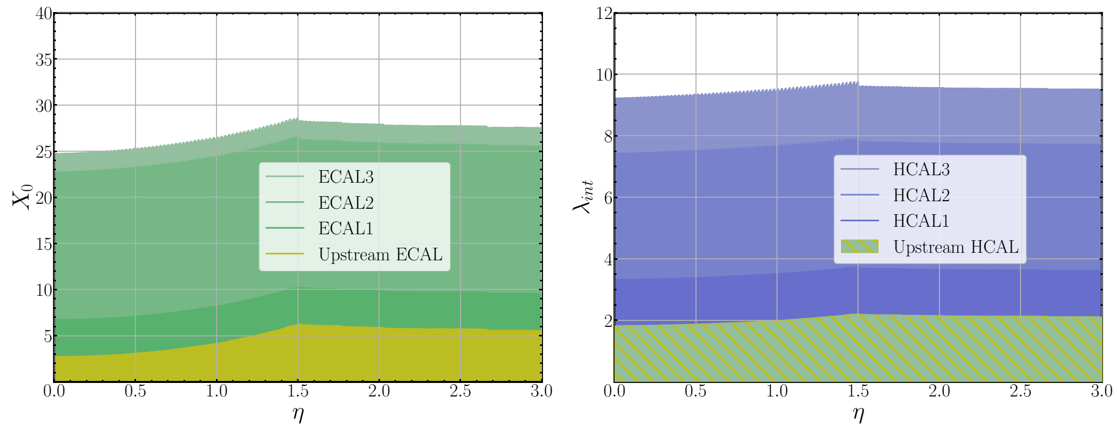


Figure 3. (left) Cumulative amounts of material, in units of radiation length X_0 and as a function of η, in front of and within the electromagnetic calorimeter system. (right) Cumulative amounts of material, in units of interaction length λ_{int} and as a function of η, in front of and within the hadronic calorimeter system.

The inner surface of the calorimeter system is a cylinder with a radius of 150 cm and a length of 6387.8 mm immediately enclosing the iron layers and the ITS. The calorimeters are separated into a central barrel region covering the pseudorapidity range $|\eta| < 1.5$ and two end-cap regions extending the coverage up to $\eta = 3$ by default. Both the ECAL and the HCAL are divided into 3 concentric layers, with each layer being further segmented into cells with edges of constant η and ϕ. The cell granularity for each layer is configurable by setting the number of equal divisions in η and (separately) ϕ. The depth of the cells in every layer is designed to be nearly constant in η to ensure that the fraction of a particle's energy deposited in each layer does not depend on the incident angle. This design, leading to layer shapes of the form $1 / \cosh\eta$, provides a uniform calorimeter thickness as a function of pseudo-rapidity. CoCoA will thus have a more uniform response than a pure circular cylindrical shape.

The CoCoA calorimeter material is a compound using an equivalent molecule approximation, mixing an absorber and an active material with a constant proportion. Both the materials and their proportion can be configured for the ECAL and the HCAL individually. By default, the ECAL is made of a mixture of lead and liquid argon, corresponding to the ATLAS ECAL materials. The volume proportion amounts to 1 : 3.83, resulting in a radiation length of $X_0 = 2.5$ cm. The ECAL and HCAL are separated by an iron layer with a default thickness of 80 mm. The HCAL is made of a mixture of iron and polyvinyl toluene plastic material with a volume proportion of 1.1 : 1.0, resulting in a nuclear interaction length of $\lambda_{\text{int}} = 26.6$ cm. The integrated radiation and interaction length measured from the interaction point (IP) at the center to the end of the HCAL is shown as a function of the pseudorapidity in figure 3.

Table 1. Calorimeter default design values regarding layer depths in terms of radiation lengths X_0 (ECAL) and hadronic interaction lengths λ_{int} (HCAL), granularity and energy noise levels.

Layer	Depth	Segmentation ($\eta \times \phi$)	Std. dev. noise per cell per event (MeV)
ECAL 1	$4\,X_0$	256×256	13
ECAL 2	$16\,X_0$	256×256	34
ECAL 3	$2\,X_0$	128×128	41
HCAL 1	$1.5\,\lambda_{\text{int}}$	64×64	75
HCAL 2	$4.1\,\lambda_{\text{int}}$	64×64	50
HCAL 3	$1.8\,\lambda_{\text{int}}$	32×32	25

While this calorimeter design represents a homogeneous detector, a spread in the resolution of reconstructed energies in accordance with a sampling calorimeter design is emulated by means of configurable sampling fraction parameters for the ECAL and the HCAL individually. In lieu of a complete simulation of active and passive material, the sampling is emulated by accounting only for a fraction of the GEANT4 energy deposition steps for all particles in the calorimeter showers. The steps to be removed are chosen randomly. The sum of the total deposited energy by those steps is computed and the total energy released is estimated by inverse scaling of the total deposited energy by the corresponding fraction.

Noise, as for example from electronics, is simulated by the addition of random amounts of energy following a Gaussian distribution centered around zero. The noise is independently added to each cell. Negative energies are allowed as is typically the case as a result from the subtraction of pedestals. If such downward fluctuations are significant in size, those negative energy cells can be clustered into topoclusters. The implementation of pile-up collision events from additional proton interactions is left for future development. The default choices of materials and smearing parameters provided in table 1 are chosen in order to approximate single-particle responses of the ATLAS calorimeter system [13, 14].

3. Data processing

Every event is processed according to the workflow presented in figure 4. First, primary particles are generated at the IP by means of the PYTHIA8 Monte Carlo event generator [15]. A broad range of primary physics processes is available to the user, ranging from the generation of single particles as well as single jets up to more complicated final states with large multiplicities of jets and leptons in the final state.

The set of final state, stable particles is stored in the output file and passed on to the detector simulation described in the previous section, where the propagation of these particles and their interactions with the detector material is simulated in GEANT4 [7]. The model of hadronic interactions is chosen in accordance with the ATLAS and CMS detector simulations. The sum of the energies deposited in each calorimeter cell is stored. Electronic noise is simulated by the addition of random energy offsets to each cell for which table 1 provides the default values of standard deviation for each layer.

For the purpose of particle reconstruction, the origin of energy deposits in each cell is stored via a list of parent particle indices which contributed energy into the cell and a list of weights recording their relative contribution to the total cell energy. Cells which received their dominant contribution from electronic noise are assigned an index of -1.

In order to limit the number of calorimeter cells stored in the output file to a reasonable level, low-energy cells dominated by noise contributions are suppressed using a topological clustering algorithm [16]. Only cells contained in the resulting 'topoclusters' are stored in the output file. Topoclusters are seeded by single cells which are required to contain a deposited energy well above the noise level, where the threshold of this signal-to-noise ratio (SNR) is 4.6 for CoCoA by default, while a value of 4.0 is used for the ATLAS experiment. This difference is chosen in order to achieve a better agreement between ATLAS and CoCoA in terms of the topocluster multiplicity distribution for single charged and neutral pions as well as pure noise events (figure 5). Starting with the seeding cells, all neighboring cells are added to the cluster if their SNR is above another threshold, where the default value is set to 2. Finally, all further neighboring cells above a third threshold are added, which by default is set to 0. Cells with negative energy can be included, based on their absolute value, or excluded entirely (default configuration). Topocluster candidates containing multiple local maxima in ECAL cell energy each surpassing 400 MeV are split into separate topoclusters.

In order to support particle reconstruction studies which include high energy primary photons, electron–positron pairs from photon conversions taking place in the ITS upstream its two outermost iron layers are stored in the CoCoA output file as well. Tracks emanating from photon conversions and also

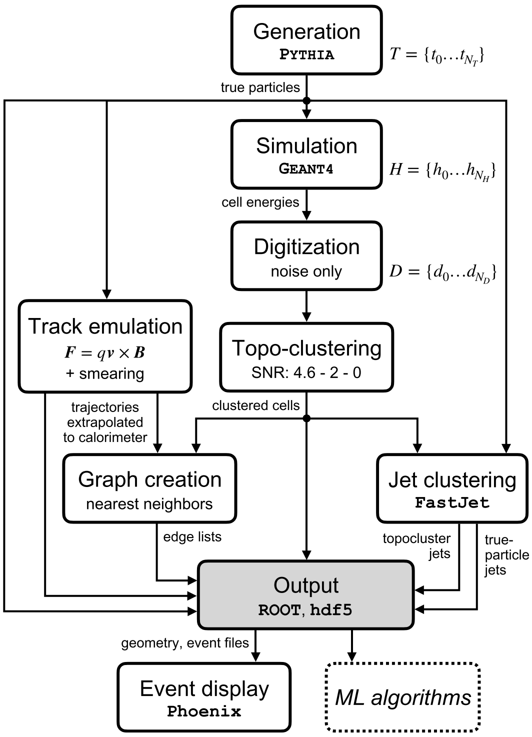


Figure 4. CoCoA workflow. Primary particles generated with the PYTHIA library are introduced to CoCoA. Their interactions with the detector material is simulated by means of the GEANT4 toolkit. Calorimeter cells identified by a topological clustering algorithm are stored in the output ROOT file together with true particles, emulated tracks, and particle trajectories extrapolated from the IP through the calorimeter according to the equations of motion. A nearest-neighbors-based graph is constructed and stored via edge lists connecting source and destination nodes amongst the output cells and tracks. Jets made of true particles as well as topoclusters are stored in the output file as well. Events in the output file can be parsed for visualization in PHOENIX.

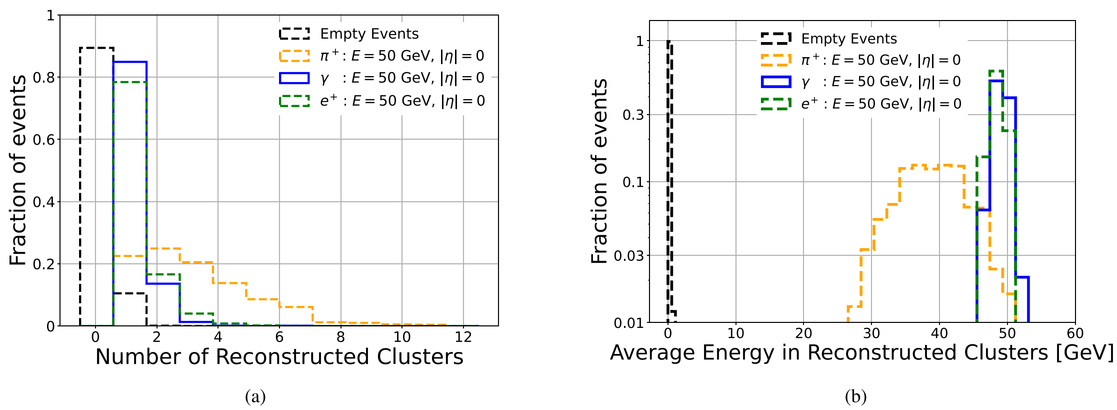


Figure 5. Number (a) and average energy (b) of reconstructed clusters in CoCoA for events with a single charged pion, electron, or photon shot at $\eta = 0$. Results are also shown for clusters reconstructed in empty events due to electronic noise. The default topological calorimeter cell clustering settings are used.

primary electron tracks are used to construct groups of topoclusters denoted as 'superclusters' associated with electron and photon showers. The superclustering procedure in CoCoA follows the criteria described in [17], designed to improve electron energy reconstruction by incorporating nearby energy deposits from bremsstrahlung. It also includes criteria for grouping multiple clusters that are related by a pair of nearby tracks to a photon conversion vertex, thus improving reconstructed photon energy. In the photon conversion shown in figure 1, for example, the CoCoA output contains a supercluster which combines the three topoclusters shown. Due to the simplified tracking, the criteria on number of track hits are not applied. The CoCoA implementation does not focus on electron and photon identification; rather, superclusters are only formed using tracks linked to primary or conversion electrons.

While the event simulation based on the GEANT4 [7] toolkit determines particle trajectories according to their equations of motion and their interactions with detector material, CoCoA implements a particle

Table 2. Default k (number of nearest-neighbors) and maximum ΔR separation used to define edges in the fixed graph creation. Edges between cells are denoted 'c–c' while edges between tracks and cells are denoted 't–c'.

	ECAL layer			HCAL layer		
	1	2	3	1	2	3
k (c–c) inter-layer	1	2	2	2	2	1
k (c–c) same layer	8	8	8	6	6	6
k (t–c)	4	4	4	3	3	3
$\Delta R^{\max}$ (c–c)	0.05	0.07	0.14	0.30	0.30	0.60
$\Delta R^{\max}$ (t–c)	0.15	0.15	0.40	1.10	1.10	2.00

tracking based only on the equations of motion for the benefit of downstream tasks. These tracks are extrapolated to the entry surface as well as each layer of the calorimeter and the resulting η and ϕ coordinates are stored in the output file.

For user convenience, an interface to the FastJet [18] library is provided that clusters primary particles as well as topological calorimeter cell clusters into jets. The user can choose the specific jet clustering algorithm accordingly, with the anti-k_{T} algorithm set as default.

For each event in the output data a fixed heterogeneous graph containing cells and tracks is provided by means of two lists storing the indices of source and destination nodes for each edge. The edges are created based on k nearest neighbors in angular distance with k being user-configurable per calorimeter layer and edge type. Three edge types are defined: track-to-cell, cell-to-cell inter-layer, and cell-to-cell across neighboring calorimeter layers (tracks are not directly connected). The user can configure for each of these types both how many edges to construct in a ΔR-ordered neighborhood and also with a maximum ΔR (where $\Delta R^2 = \Delta\eta^2 + \Delta\phi^2$). The default values are given in table 2.

The final output file produced by CoCoA stores an array of features for each event which are associated with the following sets: cells that participated in topoclusters, tracks, topoclusters, truth particles and decay record, graph edges, and jets. The output file format is `ROOT` but can be converted to `hdf5` format using a script provided in the repository.

4. Detector performance

In the following, the performance of CoCoA is investigated by means of single particles which are generated at the IP. For each particle type and momentum under investigation, the event generation is repeated in order to gather a statistically significant amount of events.

The correct reconstruction of particle energies is demonstrated in figure 5, which compares the distributions of multiplicities (figure 5(a)) and energy sums (figure 5(b)) of topoclusters for charged pions, photons, electrons and events containing only noise contributions, denoted as empty events. In most of the empty events, the cell energies do not pass the noise threshold of the clustering algorithm. For those events in which this threshold is passed, the average cluster energy sum amounts to 36 MeV in line with the low noise levels provided in table 1. The photons and electrons mostly result in one cluster, while their energy is reconstructed with only a small variation. In comparison, the charged pion events result in larger variations of the cluster multiplicity and energy sum distributions due to the higher degree of variations in deposited energies for the hadronic showers. The average cluster energy sum is below the initial charged pion energy due to the involved nuclear interactions of the shower particles with the detector material, which are not counted as detectable energy. A hadronic calibration procedure is not performed within CoCoA but left for downstream tasks.

Patterns of energy depositions across the calorimeter are demonstrated in figure 6 in terms of fractions of deposited energy per calorimeter layer for electrons, photons and charged pions. As a consequence of the material budget presented above in figure 3, the electrons and photons deposit most of their energy in the ECAL, in particular in the second calorimeter layer, while the charged pions reach the HCAL layers where they deposit most of their energy, in line with energy deposition patterns at collider-detector experiments.

Figure 7 shows distributions of the reconstructed energies for central electrons, photons and charged pions with different initial energies. The energy resolution provided by the calorimeter response improves as the initial particle energy increases. It is larger for charged pions compared to electrons and photons, as expected because of the existence of large sampling fluctuations for hadronic showers compared to electromagnetic showers.

Figure 8 shows the reconstructed energies of a single electron emitted at different initial η. The average reconstructed energy is always lower than the initial particle energies, with the difference growing with the

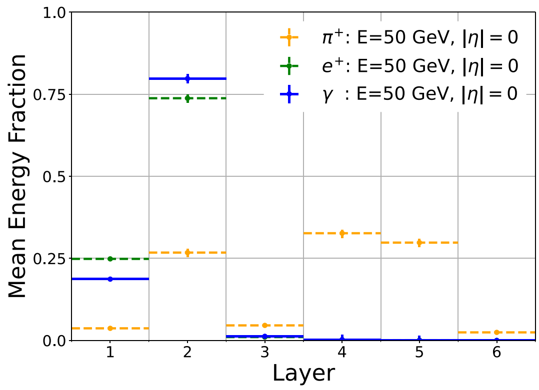
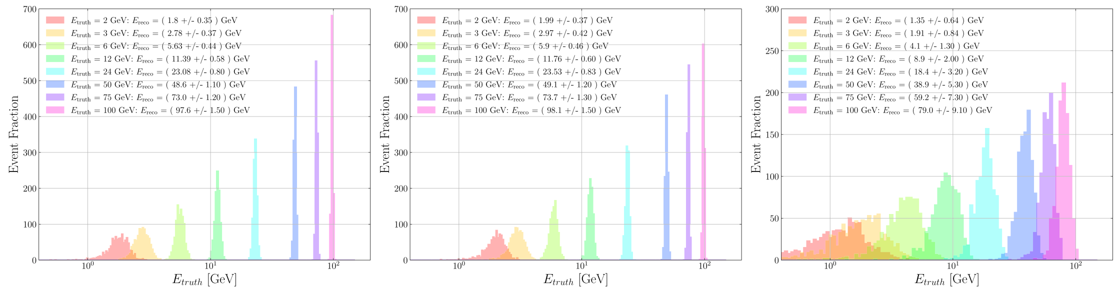


Figure 6. Energy deposited by electrons, photons and charged pions for each calorimeter layer. The electron and photon showers are limited to the electromagnetic calorimeter (layers 1–3) while the charged pion showers reach deep into the hadronic calorimeter (layers 4–6).



(a) (b) (c)

Figure 7. Reconstruction energies for central electrons (a), photons (b) and charged pions (c). The average reconstructed energies follow the initial particle energies while the energy resolution improves as the initial particle energy increases.

particle η. This is due to the energy depositions in the iron contained in the ITS upstream the calorimeter, in accordance with the material map presented in figure 3.

Figure 9 quantifies the energy resolution as a function of true particle energy, comparing electrons with charged pions. For each particle type, the relative energy resolution depending on the particle energy is fitted using least-squares to the following common form of the resolution function:

$$\frac{\sigma(E_{\text{reco}})}{E_{\text{truth}}} = \frac{a}{\sqrt{E_{\text{truth}}}} \oplus \frac{b}{E_{\text{truth}}} \oplus c \tag{1}$$

where the best-fit parameters are provided within the figure. Here a, b and c refers to the stochastic, electronic noise and constant terms, respectively. The larger fitted coefficient of the sampling term for hadronic shower compared to electromagnetic is related to the larger value of sampling fraction f configured for the ECAL and HCAL separately (0.07 and 0.025, respectively). The values of the parameters, appearing in equation (1), are individually evaluated for photon as $a = (0.16 \pm 0.01)\,\text{GeV}$, $b = (0.30 \pm 0.02)\sqrt{\text{GeV}}$ and $c = 0.006 \pm 0.003$. The same numbers for charged pions are found to be $a = (0.50 \pm 0.12)\,\text{GeV}$, $b = (0.32 \pm 0.06)\sqrt{\text{GeV}}$ and $c = 0.086 \pm 0.002$. The noise term is compatible with the input noise values, the sampling term is as expected from the sampling emulation.

The performance of the simulated detector has been so far probed using single particles. To illustrate the detector performance in a more realistic event environment, proton collision producing an on-shell W boson, decaying to an electron and neutrino, i.e. $pp \rightarrow W \rightarrow e + \nu$ were simulated. The electron is reconstructed using the superclustering algorithm described in section 3, and its energy is calibrated in order to compensate for the loss due to scattering in the ITS and iron layers upstream the ECAL. The missing transverse momentum (MET) is calculated from the rescaled clusters, as the opposite the vector sum over visible transverse momenta in the whole event. Finally the transverse W mass, m_T^W, is computed from the reconstructed W four-momentum and compared with the corresponding truth level distribution in figure 10.

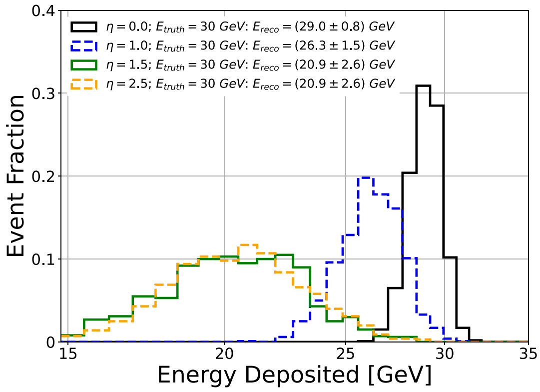


Figure 8. Reconstructed energies of electrons with an energy of 30 GeV for different directions η. As the initial electron momentum is directed closer to the beamline, the difference between the reconstructed energy and the initial particle energy increases due to the iron traversed by the electron upstream the calorimeter (dead material).

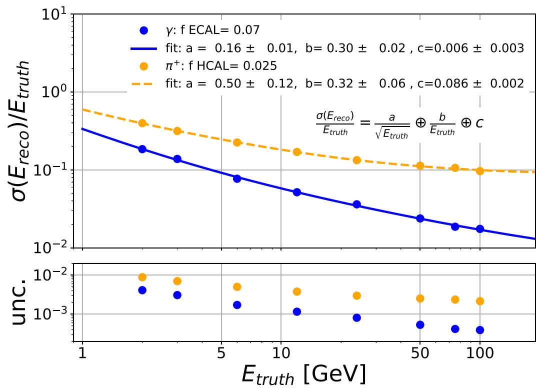


Figure 9. The relative energy resolution $\sigma(E_{\mathrm{reco}})/E_{\mathrm{truth}}$ is plotted as a function of E_{truth} for eight different truth energy values and fitted with the relative-resolution function, for photon and pion, respectively. The average sampling fraction f for ECAL and HCAL are shown in the legend.

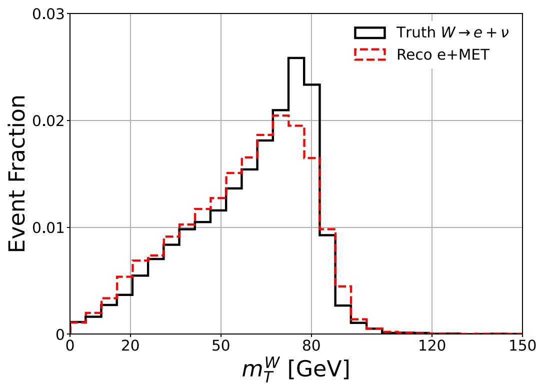
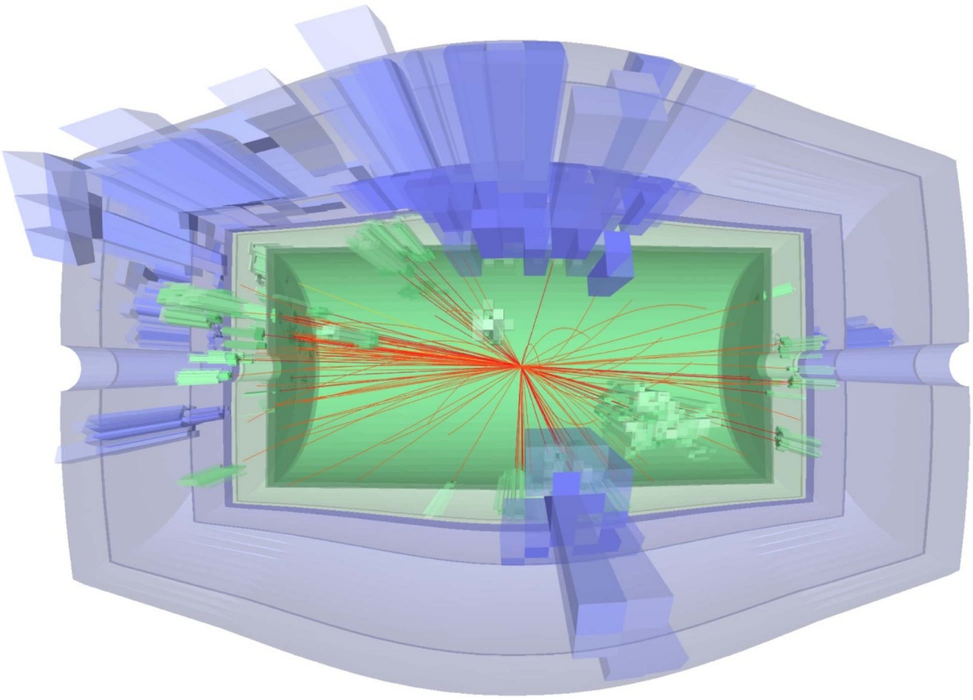


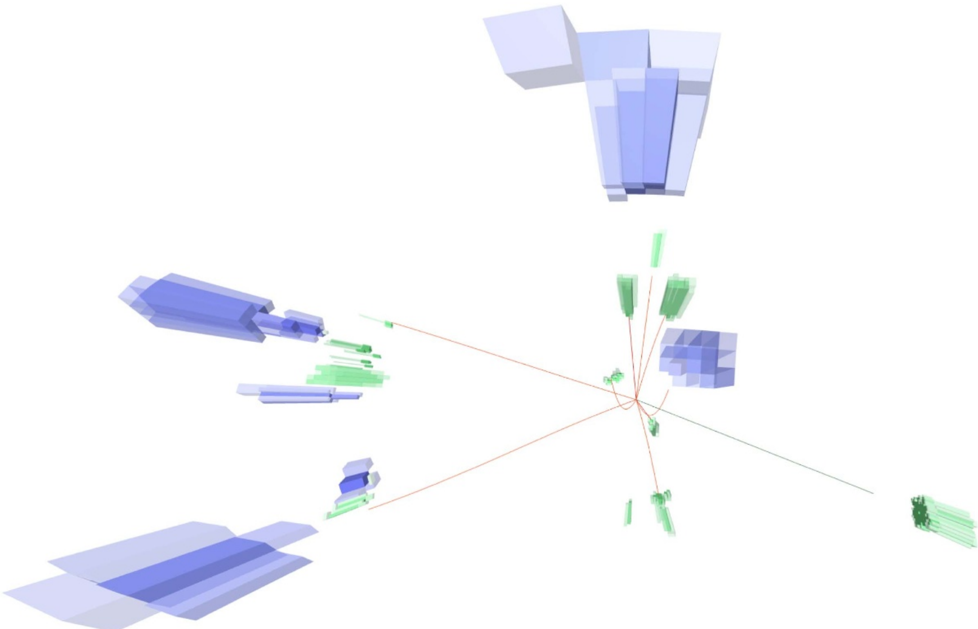
Figure 10. The transverse W mass m_T^W distribution is plotted for leptonically decaying W events. The black curve shows the truth distribution whereas the red curve is obtained from the vector sum of reconstructed lepton momentum and the MET in the event. The peak location of the two distributions are well aligned, demonstrating that the event-level reconstructed MET is trustworthy within the CoCoA framework.

5. Event display

Visualization of detector geometry and examples of hits for individual events is important for communicating results, and interpreting downstream tasks such as reconstruction and event selection. The default geometry of the CoCoA detector was ported into the open-source framework Phoenix, chosen for its versatility and user support. An example event display is shown in figure 11.



(a)



(b)

Figure 11. Phoenix event displays configured using the CoCoA detector geometry, showing the charged particle tracks and calorimeter hits generated by (a) $pp \rightarrow t\bar{t}$ and (b) $pp \rightarrow W \rightarrow e\nu$ events simulated with PYTHIA8. In (a), a cutaway of the CoCoA calorimeter volumes is shown along with the clustered cells, while in (b) only the cells are shown. The electron from the W decay in (b) is indicated by a green line. Both displays are shown in perspective view, such that nearer objects appear larger. Different shades of green and blue represent the different layers of ECAL and HCAL, respectively, while cell opacity is determined by cell SNR.

6. Conclusion

The growing interest in ML approaches to low-level analysis tasks such as event or jet reconstruction in a realistic detector underscores the importance of leveraging the rich feature space of calorimeter showers for improving these tasks. Providing an open, configurable, and realistic calorimeter simulation, CoCoA will facilitate the development of such algorithms and ultimately expand the physics reach of current and next-generation collider experiments. The thorough treatment of particle interactions in Geant4 and the full-coverage, highly-granular design of CoCoA calorimeter system enable an accurate representation of the complex data environment present in the ATLAS and CMS experiments at the LHC. To quantify this resemblance, an investigation of the single-particle response characteristics, in terms of topological clustering performance and energy resolution for electromagnetic and hadronic showers, has been carried out. Finally, additional aides including data post-processing, event visualization, and documentation for CoCoA has been provided to further encourage use.

In future, we foresee inclusion of pileup, direct interface to MadGraph [19] output file and include python binding for CoCoA.

Data availability statement

The data that support the findings of this study will be openly available following an embargo at the following URL/DOI: https://zenodo.org/record/7700475#.ZBAtZOxBzRI.

Acknowledgments

We thank Sven Menke for a careful reading of the manuscript. E G would like to thank the United States-Israel BSF, Grant NSF-BSF 2020780, for its support. E D is supported by the Zuckerman STEM Leadership Program. S G is partially supported by Institute of AI and Beyond for the University of Tokyo. K C and P R are supported by NSF award PHY-2111244, and K C is also supported by NSF award OAC-1836650. L H is supported by the Excellence Cluster ORIGINS, which is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC-2094-390783311.

ORCID iDs

Kyle Cranmer https://orcid.org/0000-0002-5769-7094
Etienne Dreyer https://orcid.org/0000-0001-8955-9510
Sanmay Ganguly https://orcid.org/0000-0003-1285-9261

References

[1] Guest D, Cranmer K and Whiteson D 2018 Deep learning and its application to LHC physics *Ann. Rev. Nucl. Part. Sci.* **68** 161–81
[2] HEP ML Community A living review of machine learning for particle physics (available at: https://iml-wg.github.io/HEPML-LivingReview/)
[3] Mertens A 2015 New features in Delphes 3 *J. Phys.: Conf. Ser.* **608** 012045
[4] Paganini M, de Oliveira L and Nachman B 2018 CaloGAN: simulating 3D high energy particle showers in multilayer electromagnetic calorimeters with generative adversarial networks *Phys. Rev.* D **97** 014021
[5] Qasim S R, Kieseler J, Iiyama Y and Pierini M 2019 Learning representations of irregular particle-detector geometry with distance-weighted graph networks *Eur. Phys. J.* C **79** 608
[6] Di Bello F A, Ganguly S, Gross E, Kado M, Pitt M, Santi L and Shlomi J 2021 Towards a computer vision particle flow *Eur. Phys. J.* C **81** 107
[7] Allison J *et al* 2016 Recent developments in Geant4 *Nucl. Instrum. Methods Phys. Res.* A **835** 186–225
[8] Ganguly S, Dreyer E, Rieck P, Heinrich L and Santi L 2023 cocoa-hep/cocoa-hep: 0.1.1 (available at: https://doi.org/10.5281/zenodo.7700475)
[9] Corentin A *et al* 2022 Open data detector (available at: https://zenodo.org/record/6445359#.Y9vBzy8RqLe)
[10] Di Bello F A *et al* 2022 Conditional generative modelling of reconstructed particles at collider experiments (available at: https://arxiv.org/abs/2211.06406)
[11] Di Bello F A *et al* 2022 Reconstructing particles in jets using set transformer and hypergraph prediction networks (arXiv:2212.01328)
[12] Moyse E, Ali F, Cortina E, Bianchi R M and Couturier B 2021 The Phoenix event display framework *EPJ Web Conf.* **251** 01007
[13] ATLAS Collaboration 2008 The ATLAS experiment at the CERN large hadron collider *J. Instrum.* **3** S08003
[14] The ATLAS collaboration 2022 Measurement of the energy response of the ATLAS calorimeter to charged pions from $W^\pm \to \tau^\pm(\to \pi^\pm \nu_\tau)\nu_\tau$ events in Run 2 data *Eur. Phys. J.* C **82** 223
[15] Bierlich C *et al* 2022 A comprehensive guide to the physics and usage of PYTHIA 8.3 *SciPost Phys. Codebases* 8 (available at: https://scipost.org/SciPostPhysCodeb.8)
[16] The ATLAS collaboration 2017 Topological cell clustering in the ATLAS calorimeters and its performance in LHC Run 1 *Eur. Phys. J.* C **77** 490

[17] The ATLAS collaboration 2019 Electron and photon performance measurements with the ATLAS detector using the 2015–2017 LHC proton-proton collision data *J. Instrum.* **14** P12006

[18] Cacciari M, Salam G P and Soyez G 2012 FastJet user manual *Eur. Phys. J.* C **72** 1896

[19] Alwall J, Frederix R, Frixione S, Hirschi V, Maltoni F, Mattelaer O, Shao H S, Stelzer T, Torrielli P and Zaro M 2014 The automated computation of tree-level and next-to-leading order differential cross sections and their matching to parton shower simulations *J. High Energy Phys.* JHEP07(2014)079