HGPflow: Particle reconstruction as hyperedge prediction

Nilotpal Kakati, Etienne Dreyer, Eilam Gross, Anna Ivina, Jonathan Shlomi
Weizmann Institute of Science
nilotpal.kakati@weizmann.ac.il
etienne.dreyer@weizmann.ac.il

Lorenzo Santi, Matteo Tusoni
Sapienza University of Rome

Francesco Armando Di Bello
University of Genova

Sanmay Ganguly
ICEPP, University of Tokyo

Lukas Heinrich
Technical University of Munich

Marumi Kado
Technical University of Munich
Sapienza University of Rome

Abstract

We approach particle reconstruction in collider experiments as a set-to-set problem and show the efficacy of a deep-learning model that predicts hypergraph incidence structure. This model outperforms a benchmark parameterized algorithm in predicting the momentum of particle jets and shows an ability to disentangle individual neutral particles in the collimated environment. Representing particles as hyperedges on the set of input nodes introduces an inductive bias that predisposes the predictions to conserve energy and thus promotes accurate, interpretable results.

1 Introduction

Inferring a set of particles based on a set of energy deposits in a detector is foundational to analysis of data from collider experiments. However, particle reconstruction is complicated by a number of factors: the high multiplicity and collimated signatures intrinsic to hadron collisions, the presence of simultaneous scattering events (pileup), and the extensive, irregular array of sensitive elements required for a highly-granular, full-coverage detector. Experiments at the Large Hadron Collider currently employ parameterized "particle-flow" algorithms which exploit complementary information provided by different detector subsystems [1, 2].

As in other applications to particle physics, deep learning (DL) brings to the particle-flow paradigm the potential to replace parameterized cuts (for example in energy subtraction schemes) with decision boundaries that leverage the full set of relevant features in data. The expressiveness of DL models also opens new possibilities such as reconstructing neutral particles inside of jets.

Previous work A proof of concept for a DL-based particle-flow algorithm was provided by [3] at the level of calorimeter cells from overlapping charged and neutral pions. Different proposals have since addressed the fuller set-to-set problem of predicting reconstructed particle candidates based on a typically much larger set of detector-level entities. In [4], the object condensation approach

Machine Learning and the Physical Sciences workshop, NeurIPS 2022.

was proposed for semi-supervised clustering of nodes in latent space to form candidate objects. This was recently applied to the task of predicting calorimeter clusters in CMS data [5, 6], where the authors focused on reconstruction efficiency and energy regression of showers from single particles embedded in pileup. The reduction in size from input to output set is handled in the MLPF [7] approach by assigning input nodes to particles in the output set or else to a dedicated "neglect class". The MLPF approach was also recently tested on CMS data [8], for the task of predicting the outputs of a traditional particle-flow algorithm.

In our hypergraph-based approach, *HGPflow*, each particle in the output set is represented by a collection (i.e. hyperedge) of the input nodes. This learned map from input to output set can be fully supervised and allows the network to exploit the relationship between target particle properties and input node features.

2 Dataset

Detector simulation Events containing a single initial state quark ($E \in [10, 200]$ GeV, pseudorapidity $\eta \in [-2.5, 2.5]$), followed by parton shower and hadronization are generated using PYTHIA8 [9]. The resulting jet of final state particles enters a tracking cylinder of radius 150 cm immersed in a uniform axial magnetic field of 3.8T wherein material interactions are emulated solely by a track momentum resolution of $\frac{\sigma(p)}{p} = p \cdot 10^{-5}/[\text{GeV}]$. A GEANT4-based calorimeter model then simulates the particles in a subsequent iron layer and 6 concentric calorimeter layers. The cylindrical layers have uniform granularity in η and azimuthal angle ϕ ranging from 256×256 in the first layer to 64×64 in the last. All final state particles with transverse momentum $p_{\text{T}} > 1$ GeV that reach the calorimeter are targets for the network.

Graph creation In each event, a topological clustering algorithm, closely resembling [10], is used to group calorimeter cells into "topoclusters" based on their proximity and deposited energy. To form an input graph, each cell (track) is connected to a maximum of 8 (4) nearest cells in the first three calorimeter layers and 6 (3) in the next three. Additionally, cells are connected to their single nearest neighbor in the immediately adjacent layer(s). Topoclusters are represented in the input graph by a separate set of nodes with edges connecting each to the set of cells belonging to the topocluster.

3 Hypergraph particle-flow network

A hypergraph is a generalization of a graph where *hyperedges* can each connect one, two, or multiple nodes (fig. 1b). The connectivity between N nodes from K hyperedges is described by an incidence matrix $I^{(N \times K)}$. In the context of particle reconstruction, calorimeter deposits and tracks are nodes in the hypergraph, while each particle is represented by a hyperedge connecting the set of nodes to which that particle contributed. We work with topoclusters rather than cells in the input set, to reduce dimensionality and to study the task of learning a non-injective map from particles to nodes.

Node encoding Fig. 1a depicts the architecture of the node encoding model used in the HGPflow network. First, two separate networks are used to embed the feature vectors of cells and tracks into a common representation space of dimension 100. Cell input features are (energy,position,η,ϕ,layer). The track input features are (p_{T},η,ϕ,d_0,z_0)[1] and the extrapolated η and ϕ coordinates of the track at each calorimeter layer. Four successive message-passing blocks are then used to update each node using its own representation, those of its neighbors, and the global representation. Following the message-passing blocks, topocluster representations are computed by the energy-weighted mean of the cell representation vectors belonging to the topocluster.

Incidence matrix prediction The first objective of the HGPflow network is to predict $(N + 1) \times K$ entries comprising a zero-padded incidence matrix and an additional row of binary values that indicates whether the particle corresponding to a given column exists or not (where $K = 30$ is an upper bound on the number of particles in the training data). We implement the recurrent training strategy of 16 refinement blocks truncated by random gradient skips

[1] d_0 and z_0 are the distance of closest approach of the track to the beam line in the transverse and longitudinal directions, respectively.

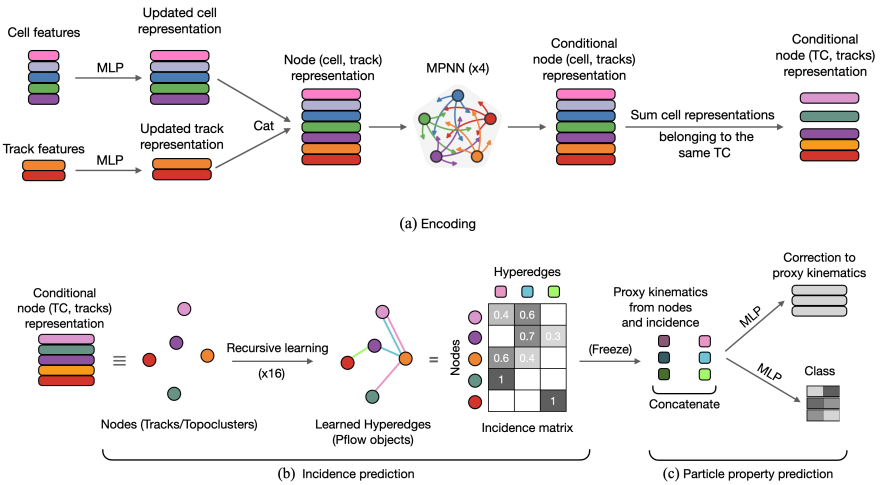


Figure 1: The three main components of the HGPflow network: (a) the encoding model (b) recurrent model for learning the incidence matrix and the indicator, and (c) particle property prediction network.

described in [11], and extend this approach to the case of a fractional rather than binary-valued incidence matrix. We define the target entry relating node i to particle a as follows:

$$[I]_{ia} = \frac{E_{ia}}{\sum_{\text{particles } b} E_{ib}} = \frac{E_{ia}}{E_i} \tag{1}$$

where E_{ia} is the amount of energy that particle a contributes to the total energy E_i of node i (simply equal to 1 for tracks). Predicted rows in the incidence matrix are normalized using Softmax (i.e. sum over all hyperedges for a given node is 1) before being compared to target via Kullback–Leibler divergence loss. Predicted columns are rearranged using the Hungarian algorithm [12] to minimize the loss. An example of target and predicted incidence matrix entries is shown in fig. 2.

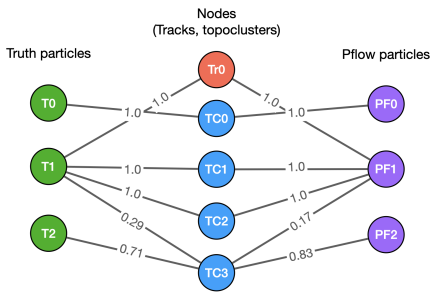


Figure 2: The truth and predicted fractional incidence matrix entries connecting track (Tr) and topocluster (TC) nodes to particles for an example from the test dataset.

Properties prediction The third component of the HGPflow network (fig. 1c) predicts particle properties for each hyperedge. Classification between photons and neutral hadrons is performed for hyperedges which do not contain a track and are thus identified as neutral particles. Predicting the incidence matrix (eq. 1) enables a unique advantage: kinematics of neutral particle can be approximated as weighted sums and averages over the input features of the topoclusters contained in the hyperedge. Proxy quantities (denoted $\hat{}$) for energy and angular coordinates can be computed as:

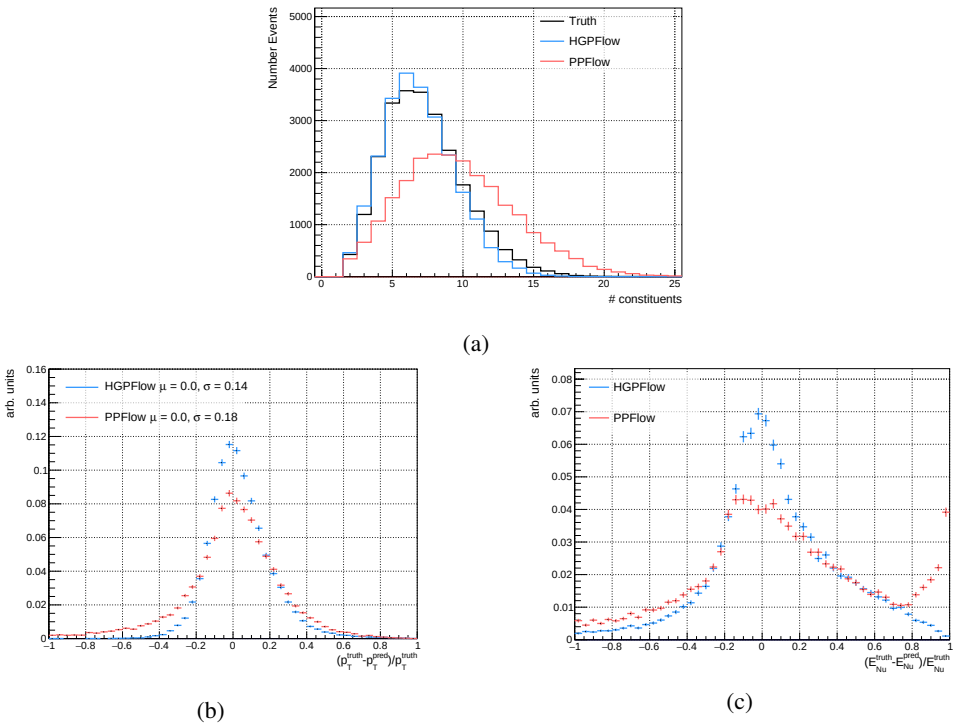
$$\hat{E}_a = \sum_{\text{nodes } i} E_i I_{ia} \,, \qquad \{\hat{\eta}_a, \hat{\phi}_a\} = \sum_{\text{nodes } i} \{\eta_i, \phi_i\} \tilde{I}_{ia} \tag{2}$$

where a dual incidence matrix $\tilde{I}$, normalized over node instead of particle indices, can be defined:

$$\tilde{I}_{ia} = \frac{E_{ia}}{\sum_{\text{nodes } j} E_{ja}} = \frac{E_{ia}}{E_a} = \frac{E_i \cdot I_{ia}}{\sum_{\text{nodes } j} (E_j \cdot I_{ja})} \tag{3}$$

Therefore, neutral particle kinematics (p_T, η, ϕ) are regressed by predicting an offset to the proxy values in eq. 2. For charged particles, an offset is likewise predicted for the p_T measured from the associated track. The properties loss is computed by matching predicted and target particles using the Hungarian algorithm. Particles corresponding to hyperedges where the predicted indicator was below threshold are matched to dummy targets and weighted by zero in the loss.

The training, validation, and test datasets for our results contain 50000, 5000, and 32328 single-jet events, respectively, following the description in section 2. The model is trained for roughly 4 days on an 24564MiB GPU (NVIDIA RTX A5000), completing 65 + 15 epochs over which the consecutive trainings of incidence and particle properties converge. Data and code will be made available at the following repository: https://github.com/nilotpal09/hg-tspn-pflow



(a) (b) (c)

Figure 3: (a) Number of particles (or topoclusters and tracks in PPFlow case) clustered in the jet. Distribution of relative residuals between predicted and true (b) jet p_{T} and (c) total energy from neutral particles per event for both HGPflow and PPflow.

4 Performance of particle reconstruction in jets

After evaluation on the test dataset, target charged particles are paired with predicted charged particles based on an angular match between the associated track. Neutral target and predicted particles are paired using the Hungarian algorithm with distance metric $\sqrt{(\Delta p_T/p_T^{\mathrm{true}})^2 + \Delta\eta^2 + \Delta\phi^2}$.

Particle-level performance Using calorimeter information enables the HGPflow network to predict p_{T} for charged particles with a resolution that is 9.8% better than the track measurement at 15 GeV and 36% better above 40 GeV. Table 1 summarizes the efficiency ($N_{\mathrm{matched}}/N_{\mathrm{true}}$) and fake rate ($N_{\mathrm{!matched}}/N_{\mathrm{pred}}$) of reconstructing neutral particles in the jet, and the accuracy of classifying them as either photons or neutral hadrons.

p_{T}	$1-10$ GeV	>10 GeV
Efficiency [%]	80.0	90.2
Fake rate [%]	16.8	5.0
Accuracy [%]	90.4 (52.5)	94.5 (70.0)

Table 1: Efficiency and fake rate of reconstructing neutral particles, and classification accuracy for photons (neutral hadrons).

Jet-level performance To evaluate the HGPflow performance on jet quantities, target and predicted particles are clustered using the anti-k$_t$ algorithm [13] with radius parameter 0.4.Jets are also formed using the outputs of a baseline parameterized particle-flow algorithm (PPflow) following [1]. A simple calibration is applied to center the jet p_{T} distributions for HGPflow and PPflow around zero.

The number of jet constituents is shown in fig. 3a for the three collections. The shifted PPflow distribution reflects the fact that, in contrast to HGPflow, this approach doesn't predict individual

neutral particles, but simply the fraction of neutral energy per topocluster. Although HGPflow is not trained to regress jet properties directly, it outperform the PPflow baseline in reconstructing the jet p_{T} (fig. 3b) and the total energy from neutral particles per event (fig. 3c). The peak at 1 in fig. 3c for PPflow results when the parameterized subtraction removes all energy present. An example event display showing reconstructed particles in a jet is provided in the appendix.

5 Discussion and summary

Inductive bias The jet-level performance of HGPflow can be understood in terms of its prediction of charged particle p_{T} that improves the track measurement and its ability to both reconstruct neutral particles and regress their momentum (shown in fig. 3a and 3c, respectively). The latter benefits from the fact that successfully predicting an incidence matrix defined via eq. 1 and the hyperedge indicator row entails knowing the energy contributions a given topocluster received from all particles (fig. 2). Since both the hyperedge representation and the proxy for neutral particle energy (eq. 2) are weighted by entries of the incidence matrix, the prediction which stem from them inherit a bias towards energy conservation.

Future work Applying HGPflow to full collision events including pileup will be an important extension which we expect to be straightforward. Improvements to the algorithm itself are foreseeable: first, the input graph granularity can be increased to further enable segmentation of overlapping particle showers. A second improvement could be to train the incidence and properties predictions simultaneously in a scheme that allows the two objectives to be synergistic, rather than separate.

In summary, we find that the formalism of particle reconstruction as a task of predicting hyperedges on the input set and their properties not only shows promising performance in the jet environment but also enables an interpretation of results that directly relates energy deposits to particles.

6 Broader impact

Hypergraphs have been used to model relationships on social networks (with hyperedges connecting groups of people with e.g. common friendships, publications, or product interest) [14]. We expect that in such contexts similar cases exist where not only the incidence structure of the hypergraph but also the attributes associated with its hyperedges are of interest. Moreover, our approach is particularly relevant to cases where hyperedge attributes are approximated by weighted averages over node features. However, though a similar approach to ours may help improve predictions in such cases, we do not foresee this having a negative societal impact.

7 Acknowledgments

ED is supported by the Zuckerman STEM Leadership Program. SG is partially supported by Institute of AI and Beyond for the University of Tokyo. EG is supported by the Israel Science Foundation (ISF), Grant No. 2871/19 Centers of Excellence. LH 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.

References

[1] Morad Aaboud et al. Jet reconstruction and performance using particle flow with the ATLAS Detector. *Eur. Phys. J. C*, 77(7):466, 2017.

[2] Albert M Sirunyan, CMS collaboration, et al. Particle-flow reconstruction and global event description with the CMS detector. *JINST*, 12(10):P10003, 2017.

[3] Francesco Armando Di Bello, Sanmay Ganguly, Eilam Gross, Marumi Kado, Michael Pitt, Lorenzo Santi, and Jonathan Shlomi. Towards a Computer Vision Particle Flow. *Eur. Phys. J. C*, 81(2):107, 2021.

[4] Jan Kieseler. Object condensation: one-stage grid-free multi-object reconstruction in physics detectors, graph, and image data. *The European Physical Journal C*, 80(9), Sep 2020.

[5] Shah Rukh Qasim, Kenneth Long, Jan Kieseler, Maurizio Pierini, and Raheel Nawaz. Multi-particle reconstruction in the High Granularity Calorimeter using object condensation and graph neural networks. 251:03072, 2021.

[6] Shah Rukh Qasim, Nadezda Chernyavskaya, Jan Kieseler, Kenneth Long, Oleksandr Viazlo, Maurizio Pierini, and Raheel Nawaz. End-to-end multi-particle reconstruction in high occupancy imaging calorimeters with graph neural networks. *arXiv preprint arXiv:2204.01681*, 2022.

[7] Joosep Pata, Javier Duarte, Jean-Roch Vlimant, Maurizio Pierini, and Maria Spiropulu. Mlpf: Efficient machine-learned particle-flow reconstruction using graph neural networks. *arXiv preprint arXiv:2101.08578*, 2021.

[8] Joosep Pata, Javier Duarte, Farouk Mokhtar, Eric Wulff, Jieun Yoo, Jean-Roch Vlimant, Maurizio Pierini, and Maria Girone. Machine Learning for Particle Flow Reconstruction at CMS. In *20th International Workshop on Advanced Computing and Analysis Techniques in Physics Research: AI Decoded - Towards Sustainable, Diverse, Performant and Effective Scientific Computing*, 3 2022.

[9] Torbjorn Sjöstrand, Stephen Mrenna, and Peter Z. Skands. A Brief Introduction to PYTHIA 8.1. *Comput. Phys. Commun.*, 178:852–867, 2008.

[10] Georges Aad et al. Topological cell clustering in the ATLAS calorimeters and its performance in LHC Run 1. *Eur. Phys. J. C*, 77:490, 2017.

[11] David W Zhang, Gertjan J Burghouts, and Cees GM Snoek. Recurrently predicting hypergraphs. *arXiv preprint arXiv:2106.13919*, 2021.

[12] Harold W Kuhn. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[13] Matteo Cacciari, Gavin P Salam, and Gregory Soyez. The anti-kt jet clustering algorithm. *Journal of High Energy Physics*, 2008(04):063, 2008.

[14] Jianming Zhu, Junlei Zhu, Smita Ghosh, Weili Wu, and Jing Yuan. Social influence maximization in hypergraph in social networks. *IEEE Transactions on Network Science and Engineering*, 6(4):801–811, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] Table 1 and fig. 3c demonstrate that HGPflow can reconstruct individual neutral particles inside of jets, and we explain in section 5 why we believe our approach of predicting hyperedges is key to the performance.
   (b) Did you describe the limitations of your work? [Yes] In section 2 we state a simplification of our dataset ("emulated solely"), and in 5 we phrase our limitations (e.g. partial event, no pileup) and untried ideas as opportunities for future work.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Section 6
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] But we are preparing the code to be more user-friendly and will release together with the dataset on the timescale of NeurIPS.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Section 3
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix

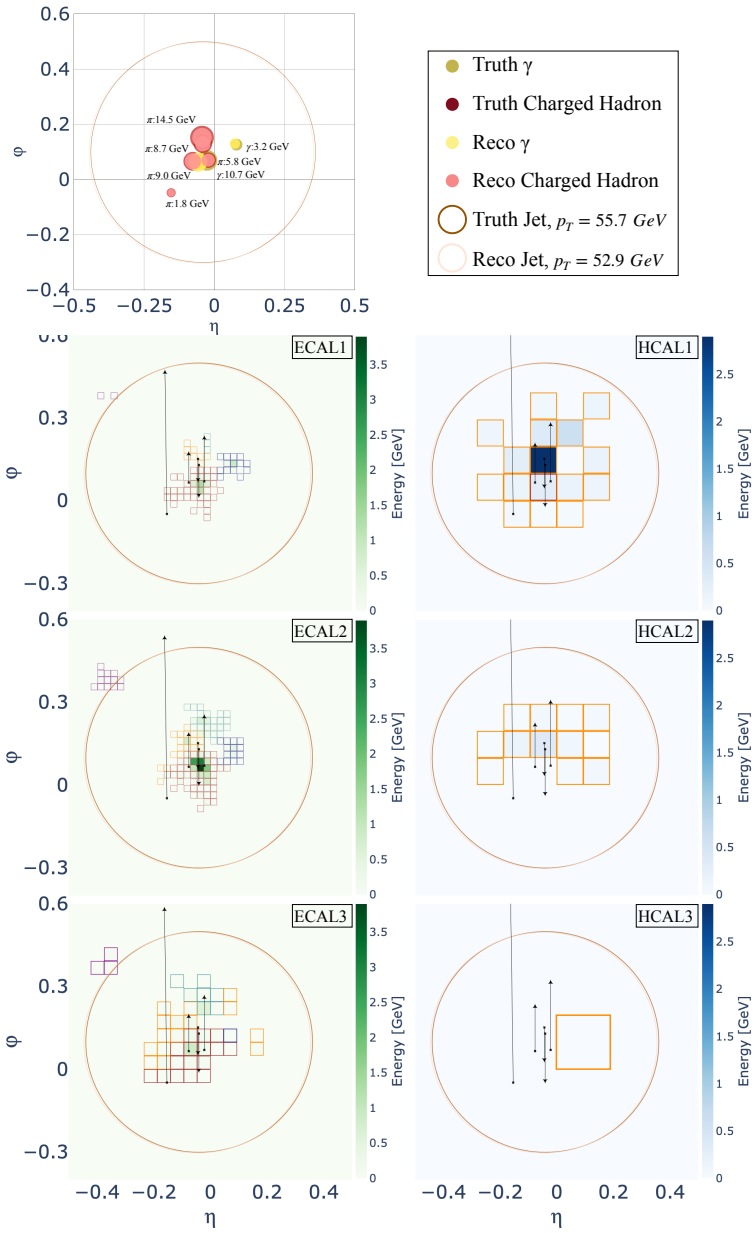


Figure 4: Event display of a single-jet event including angular locations and momenta of truth and reconstructed (predicted) particles. Each of the six calorimeter layers is shown in the region of interest. The fill of calorimeter cells shows their energy and the color of their outlines show the topocluster they belong to. The jet is composed of two photons and four charged particles, all of which are correctly reconstructed by the HGPflow network in this example, taken from the test dataset. The particles are presented at the interaction point and the tracks show their extrapolated position to the corresponding calorimeter layer. The circle of radius 0.4 represent the jet cone.