Set-Conditional Set Generation for Particle Physics

Nathalie Soybelman, Nilotpal Kakati, Etienne Dreyer, Eilam Gross, Jonathan Shlomi
Weizmann Institute of Science
nathalie.soybelman@weizmann.ac.il
nilotpal.kakati@weizmann.ac.il

Lukas Heinrich
Technical University of Munich

Sanmay Ganguly
ICEPP, University of Tokyo

Francesco Armando Di Bello
University of Genova

Marumi Kado
Technical University of Munich
Sapienza University of Rome

Abstract

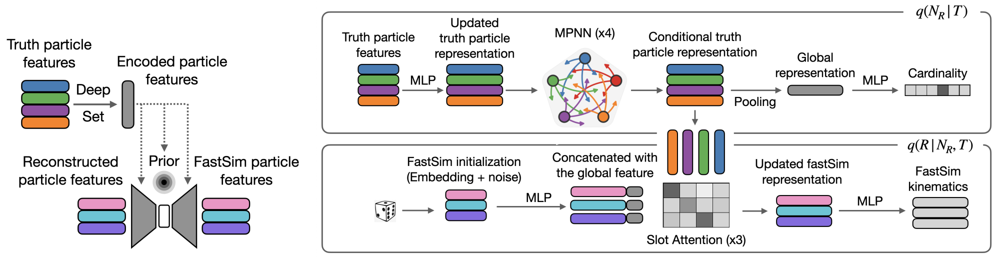
The simulation of particle physics data is a fundamental but computationally intensive ingredient for physics analysis at the Large Hadron Collider, where observational set-valued data is generated conditional on a set of incoming particles. To accelerate this task, we present an novel generative model based on graph neural network and slot-attention components, which exceeds the performance of pre-existing baselines.

1 Introduction

One of the most computationally expensive tasks in high energy physics at collider experiments is the simulation and reconstruction of collision events. Simulation tools such as Geant4 [1] use microphysical models to simulate the detailed stochastic interactions of a variable-sized *set* of "truth" particles $T = \{t_i | i = 1 \dots N_T\}$ with the detector material producing a series of signals ("hits") H in the read-out sensors of the detector. This simulation corresponds to an underlying distribution $p_{\mathrm{sim}}(H|T)$. As modern detectors have up to a hundred million of such sensors, the high-dimensional space of hits is unusable for physics analysis. "Reconstruction" is a deterministic inference algorithm $R(H)$ that attempts to recover approximately the set-valued latent input T in order to present physicists with an interpretable and low-dimensional summary of the hit data as a set of *reconstructed* particles $R = \{r_i |\ i = 1 \dots N_R\}$ that aims to approximate T. Typically, physicists do not directly interact with the hit-level data, but only with the set-valued model $R \sim p(R|T) = \int \mathrm{d}H\, \delta(R(H) - R) p_{\mathrm{sim}}(H|T)$. Therefore there is considerable interest in exploiting generative machine learning to develop fast surrogates for this effective model. In this work we explore the possibility to train an end-to-end surrogate $R \sim q_\theta(R|T)$ with learnable parameters θ. We split the generative process into a cardinality prediction task $q_{\theta_1}(N_R|T)$ and a doubly conditional set generation task $q_{\theta_2}(R|N_R, T)$. The necessary permutation invariances implied by the set nature of T and R are enforced through inductive biases in the architecture. Our results exceed the performance of baseline models.

Related Work There are two main approaches to approximate $p(R|T)$. In one approach only $p_{\mathrm{sim}}(H|T)$ is replaced by a fast surrogate such as CaloGAN [2]. Based on its output, the standard reconstruction algorithm may be used to produce reconstructed particle sets $R(H)$. This approach has two disadvantages: Firstly, the surrogate must correctly learn a high-dimensional generative model

Machine Learning and the Physical Sciences workshop, NeurIPS 2022.



(a) cVAE (b) GNN+SA

Figure 1: Architectures of the two models

$H \sim p(H|T)$ only for H to be further processed and reduced in dimension to form reconstructed events. Secondly, this approach incurs the full cost of the standard reconstruction, which is still significant. The second approach aims for an direct approximation of the lower-dimensional $p(R|T)$. Prior work simplifies the problem by first projecting the sets to fixed-sized feature vectors of the truth and reconstructed events $\mathbf{t} = f_t(T)$, $\mathbf{r} = f_r(R)$, and aims to learn a fast generative model $p(\mathbf{r}|\mathbf{t})$ [3, 4, 5]. This approach is limited by the *fixed* choice of f_r, f_t and thus does not enable one to generate features outside of f_r. A fully general approach modelling $p(R|T)$ is only possible through a *set-to-set* approach. Prior work has aimed at learning a fast surrogate using a set-valued variational autoencoder (VAE) [6], similar to the baseline we present in this work. The encoder yields a latent distribution $p(z|T)$, and a decoder implements a model of reconstructed events $p(R|z)$. While this model successfully reproduces the *marginal* distributions (i.e. projections of $p(R) = \int \mathrm{d}T\ p(R|T)p(T)$), the authors note that "[the algorithm] fails in faithfully describing the jet dynamics at constituents level" and do not present non-marginal results. To the best of our knowledge this is the first work presenting extensive analysis conditional distribution of such a set-to-set model in a particle physics application.

2 Datasets

We demonstrate the set-to-set approach using a simplified ground truth model of $p(R|T)$ for charged elementary particles represented by their direction and momentum features $(p_{\text{trans.}}, \eta, \phi)$ [1]. Truth particles are dropped deterministically as a function of their truth features t_i. The remaining ones are transformed into reconstructed particle by adding noise according to a known distribution $p(r_i|t_i)$. We generate one hundred reconstructions ("replicas") for a given truth event $R \sim p(R|T)$. The cardinalities of reconstructed objects N_R then serve as labels for the supervised training of the cardinality prediction, while the conditional empirical distribution of replicas and the truth event serve as training samples for the unsupervised generative set generation task. The training, validation, and test datasets for our results contain 2915, 500, and 3990 events, respectively. For training (evaluation) we use 25 (100) replicas per truth event.

3 Models and Training

We compare a novel neural architecture based on graph neural network and slot-attention components (GNN+SA) to a baseline model in the form of a conditional variational auto-encoder (cVAE).

Conditional VAE For the cVAE [7] we extend the typical construction of VAEs [8] by conditioning the prior $q_P(z|T)$, an encoder $q_E(z|R,T)$ and a decoder $q_D(R|z,T)$ on the truth event T. As both R and T are sets, we encode them as Deep Sets [9] before passing them into encoder and prior to ensure permutation invariance. To handle the variable number of input particles, the cVAE input is zero-padded to a maximum cardinality and the output of the decoder network includes an additional presence variable, to indicate whether the corresponding vector is to be considered a member of the output set. The threshold value for the presence variable was optimized through a grid search to 0.6. A sketch of the cVAE architecture is shown in Figure 1a.

[1] $p_{\text{trans.}}$ is the momentum transverse to the beam axis, η is related to the azimuthal angle and ϕ the polar angle

To generate reconstructed events R, a latent code is sampled from the conditional prior $z \sim q_P(z|T)$ which is then passed through the decoder to produce candidate output vectors. The presence of each particle is then sampled according to its indicator variable.

Graph Neural Network and Slot-Attention Model In addition to the cVAE, we present results on a novel model with the architecture shown in Figure 1b. In a first step a graph neural network is used to encode the conditioning set T into a high-dimensional vector representations $\{t'_i\}$. A permutation invariant pooling of these vectors is then used as an input to a fully connected network to predict a conditional density $q_{\theta_1}(N_R|T)$ for the cardinality of the output set. Once N_R is sampled, we generate the reconstructed event through a slot-attention based network that transforms noise inputs $\{\epsilon_j\}$ and particle embeddings $\{t'_i\}$ into a set of reconstructed particles $\{r_j\} = f(\{\epsilon_j\}, \{t'_i\}|N_R)$. The architecture thus *implicitly* encodes $q_{\theta_2}(R|N_R, T)$ and we can sample reconstructed events even if an explicit evaluation of the likelihood is not possible. The output generation then proceeds through a Slot Attention layer [10], in which the N_R reconstructed particle vectors are slots that attend over the provided truth particle embeddings through multiple rounds of iterative refinement. As an attention mechanism standard dot-product attention with query, key and value vectors is used to update the reconstruction slots. The slot-attention mechanism is permutation-equivariant and thus if the initial noise model is permutation invariant the implicit model will be as well.

Training The cVAE is trained on the negative evidence lower bound loss (ELBO) averaged over reconstructions R and conditioning values T. As the reconstruction term in the ELBO, $\mathbb{E}_q[q_D(R|z,T)]$, we use the permutation-invariant sum of squared distances between truth and reconstructed particles in feature space $(p_{\text{trans.}}, \eta, \phi)$ after an assignment through the Hungarian Algorithm [11]. We train for 500 epochs (7 hours).

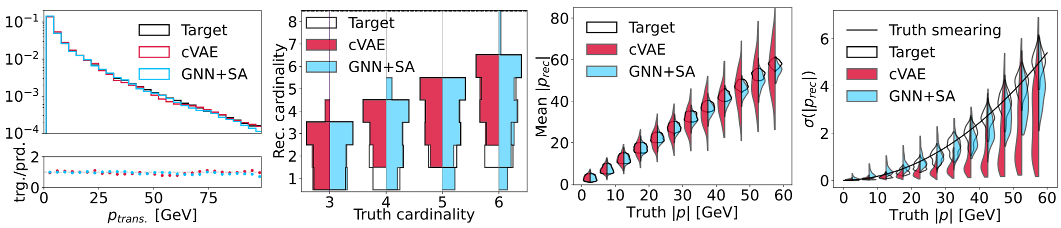
The GNN+SA model is trained on a combination of two tasks: cardinality prediction and set generation. The cardinality prediction is trained on a standard categorical cross-entropy loss $L_{\text{card.}}$ in expectation over all truth events T. As the model does not provide a tractable likelihood $q_\phi(R|T,N)$, we formulate a sample-based similarity measure between the two distributions $q_\phi(R|T,N)$ and $p(R|T,N)$. A suitable metric is the maximum mean discrepancy (MMD^2) [12] which is based on kernel functions $k(x,x')$, which act as a similarity measure between instances. We use the Hungarian Cost as the similarity measure.

$$\text{MMD}^2 = \mathbb{E}_{(x\sim p, x'\sim p)}[k(x,x')] + \mathbb{E}_{(x\sim q, x'\sim q)}[k(x,x')] - 2\mathbb{E}_{(x\sim q, x'\sim p)}[k(x,x')] \tag{1}$$

While the MMD metric enjoys strong theoretical guarantees, such as vanishing when $p = q$, we observed empirically that training directly on it as a loss converges poorly. We thus use a heuristic proxy loss which facilitates training and empirically correlates well with the MMD, which we track during training as a metric. In this proxy, we use the minimum kernel entry $L_{\text{proxy}} = \min_{x_i, x'_j} k(x_i, x'_j)$, where x_i is a member of the reference set and x'_j is from the generative model sampling. It is an upper bound on the final $-2\mathbb{E}_{p,q}[k(x,x')]$ term in the MMD definition. With these losses training on the total loss $L = L_{\text{card.}} + L_{\text{proxy}}$ is averaged over all truth events T. Due to the expensive loss we only train for 200 epochs (6 days). The models are trained on an 24564MiB GPU (NVIDIA RTX A5000). Both models are jointly optimized using the Adam optimizer [13].

4 Results

We present results for the conditional generation of reconstructed events of charged particles both in terms of per-particle features as well as collective per-event set-level features. As the ground truth model has a deterministic relationship between the truth event T and output cardinality N_R, we can assess the correctness of the model $p(N|T)$ through comparison of the accuracy of the cardinality prediction with the ground-truth cardinality. As shown in Table 1 both models perform similarly after tuning cVAE hyper-parameters 1. We also observe that marginal distributions are comparably reproduced by both models confirming prior work in this area. Figure 2a shows the per-particle momentum as an example. Differences in the two models begin to emerge when studying projections of the *conditional* distributions $p(R|T)$. We present projected distributions $p(\mathbf{r}|\mathbf{t}) = p(f_r(R) = \mathbf{r}|f_t(T) = \mathbf{t})$, where $f_t(\cdot), f_r(\cdot)$ extract feature vectors on T and R respectively.



(a) $p(p_{\mathrm{trans.}})$ (b) $p(N_R|N_T)$ (c) $p(\overline{(|p|_R)}||p|_T)$ (d) $p(\sigma(|p|_R)||p|_T)$

Figure 2: (a) Marginal $p_{\mathrm{trans.}}$ distribution, (b) Conditional Cardinality Distribution and (c and d) Distribution of reconstructed particle momentum means and their variances

	Accuracy $q(N_R|T)$ [%]	MMD^2 $q(R|N,T)$	Hungarian Cost $C(R,T)$ $|\overline{C}_q - \overline{C}_p|$
GNN+SA	$\mathbf{81.2 \pm 0.2}$	$\mathbf{-0.004 \pm 0.029}$	$\mathbf{0.026}$
cVAE	80.5 ± 0.2	0.037 ± 0.037	0.089

Table 1: Performace metrics comparing cardinality prediction accuracy, the MMD metric and the difference in the mean Hungarian Cost between reconstructed and truth events of the surrogate models to the ground truth.

In Figure 2b, we compare the learned cardinality distributions as a function of the truth cardinality $p(N_R|N_T)$. Both models broadly reproduce the target, however the GNN+SA achieves better modelling at lower cardinalities.

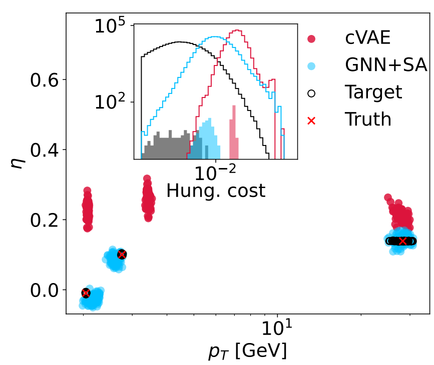


Figure 3: Event display showing the improved performance of GNN+SA.

In Figure 2c we compare the per-particle mean reconstructed momentum in bins of truth-momentum. In the ground truth reconstruction, the distribution of reconstructed momenta is as expected centered around the true momenta, with the width reflecting the variance of the noise model and a residual variance contributed to the finite size bin-width in the conditional feature $\mathbf{t}$. Here we can see that while the GNN+SA model does not fully match the ground truth it performs markedly better than the cVAE model. The high variance of the cVAE model indicates that it does not model the mean of the reconstructed particle distributions correctly. In Figure 2d the same analysis is performed for the variance of the reconstructed feature distribution and compared to the underlying ground-truth noise model that was applied to the truth particles. Here, the difference between the two models, becomes even more apparent: Whereas the GNN+SA model does track the ground truth model, albeit with a degree of under-estimation and increased variance at high momenta, the cVAE model does not manage to correctly capture the momentum dependence of the resolution to a satisfying degree. The failures of the cVAE model are apparent in the representative truth event shown in Figure 3. Comparing the results in Figure 2a and Figures 2b to 2d underlines the importance of a detailed study of the learned set-valued distribution – which is first done in this work – as mismodeling may not be apparent from marginal distributions alone. Finally, we present a metric that aims to distill the interplay between multi-particle correlation as well as the shifts in mean and variance modelling observed in the previous section into a single number. In the inset of Figure 3 we show the distribution of the Hungarian Loss $C(R,T)$ of the reconstructed events to the truth events as averaged over the full test set. The mean GNN+SA cost lies much closer to the mean ground-truth cost as compared to the cVAE model. To give a sense of scale for the significant improvement a single event is shown in the main pane of Figure 3. The cVAE fails to sample reconstructed particles correctly yielding a high Hungarian Cost. The GNN+SA samples resemble the target to a markedly higher degree. The cost distributions for this single truth event are shown as filled histograms in the inset. They each lie in the bulk of the truth-averaged distribution. The shown event is thus a representative example of the model performance. Similarly, the GNN+SA exceeds the cVAE performance as measured by MMD^2 metric as listed in Table 1.

5 Conclusions

We have presented an approach for a set-conditional set generation model to approximate simulation and subsequent reconstruction. We split the task into two-step generative procedure of cardinality prediction followed by conditional set generation and choose appropriate permutation-invariant architectures through message-passing graph neural networks and slot-attention (GNN+SA). Results are shown on the reconstructions of local collection of noised truth particles and compared to a baseline model that uses a cVAE architecture. The GNN+SA model outperforms the baseline model and better captures key properties of the target distribution. Although the results are not yet suitable for a real physics application, they are a significant improvement over prior state of the art.

6 Broader Impacts

Research on generative modelling has the potential to be misused for generating harmful content. However this work focuses on applications in particle physics and thus no negative societal impacts are expected.

References

[1] S. Agostinelli et al. GEANT4: A simulation toolkit. *Nucl. Instrum. Meth.*, A506:250–303, 2003.

[2] Michela Paganini, Luke de Oliveira, and Benjamin Nachman. CaloGAN : Simulating 3D high energy particle showers in multilayer electromagnetic calorimeters with generative adversarial networks. *Phys. Rev. D*, 97(1):014021, 2018.

[3] Georges Aad et al. AtlFast3: the next generation of fast simulation in ATLAS. *Comput. Softw. Big Sci.*, 6:7, 2022.

[4] Anja Butter, Tilman Plehn, and Ramon Winterhalder. How to GAN LHC Events. *SciPost Phys.*, 7(6):075, 2019.

[5] Raghav Kansal, Javier Duarte, Breno Orzari, Thiago Tomei, Maurizio Pierini, Mary Touranakou, Jean-Roch Vlimant, and Dimitrios Gunopulos. Graph generative adversarial networks for sparse data generation in high energy physics, 2020.

[6] Mary Touranakou, Nadezda Chernyavskaya, Javier Duarte, Dimitrios Gunopulos, Raghav Kansal, Breno Orzari, Maurizio Pierini, Thiago Tomei, and Jean-Roch Vlimant. Particle-based fast jet simulation at the LHC with variational autoencoders. *Mach. Learn. Sci. Tech.*, 3(3):035003, 2022.

[7] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[8] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2014.

[9] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. *CoRR*, abs/1703.06114, 2017.

[10] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-Centric Learning with Slot Attention. *arXiv e-prints*, page arXiv:2006.15055, June 2020.

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

[12] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13(null):723–773, mar 2012.

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (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] The code and the data will be released in the time frame of the workshop.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (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]