Zephyr quantum-assisted hierarchical Calo4pQVAE for particle-calorimeter interactions

Ian Lu
TRIUMF,
Vancouver, BC V6T 2A3, Canada
ilu@triumf.ca

Hao Jia
Department of Physics and Astronomy,
University of British Columbia,
Vancouver, BC V6T 1Z1, Canada
haojia@phas.ubc.ca

Sebastian Gonzalez
TRIUMF,
Vancouver, BC V6T 2A3, Canada
bastigonzalez2000@gmail.com

Deniz Sogutlu
TRIUMF,
Vancouver, BC V6T 2A3, Canada
dsogutlu@triumf.ca

J. Quetzalcoatl Toledo-Marin
TRIUMF,
Vancouver, BC V6T 2A3, Canada
Perimeter Institute for Theoretical Physics,
Waterloo, ON, N2L 2Y5, Canada
jtoledo@triumf.ca

Sehmimul Hoque
University of Waterloo,
ON, N2L 3G1, Canada
s4hoque@uwaterloo.ca

Abhishek Abhishek
Department of Electrical
and Computer Engineering,
University of British Columbia,
Vancouver, BC V6T 1Z4, Canada
abhiabhi@student.ubc.ca

Colin Gay
Department of Physics and Astronomy,
University of British Columbia,
Vancouver, BC V6T 1Z1, Canada
cgay@phas.ubc.ca

Roger Melko
Perimeter Institute
for Theoretical Physics,
Waterloo, ON,
N2L 2Y5, Canada
rgmelko@uwaterloo.ca

Eric Paquet
Digital Technologies Research
Centre, National Research Council,
1200 Montreal Road,
Building M-50 Ottawa,
ON, K1A 0R6, Canada
eric.paquet@nrc-cnrc.gc.ca

Geoffrey Fox
University of Virginia,
Computer Science and
Biocomplexity Institute,
994 Research Park Blvd,
Charlottesville,
VA, 22911, USA
gcfexchange@gmail.com

Maximilian Swiatlowski
TRIUMF,
Vancouver, BC V6T 2A3, Canada
mswiatlowski@triumf.ca

Wojciech Fedorko
TRIUMF,
Vancouver, BC V6T 2A3, Canada
wfedorko@triumf.ca

Abstract

With the approach of the High Luminosity Large Hadron Collider (HL-LHC) era set to begin particle collisions by the end of this decade, it is evident that the computational demands of traditional collision simulation methods are becom-

Machine Learning and the Physical Sciences Workshop, NeurIPS 2024.

ing increasingly unsustainable. Existing approaches, which rely heavily on first-principles Monte Carlo simulations for modeling event showers in calorimeters, are projected to require millions of CPU-years annually—far exceeding current computational capacities. This bottleneck presents an exciting opportunity for advancements in computational physics by integrating deep generative models with quantum simulations. We propose a quantum-assisted hierarchical deep generative surrogate founded on a variational autoencoder (VAE) in combination with an energy conditioned restricted Boltzmann machine (RBM) embedded in the model's latent space as a prior. By mapping the topology of D-Wave's Zephyr quantum annealer (QA) into the nodes and couplings of a 4-partite RBM, we leverage quantum simulation to accelerate our shower generation times significantly. To evaluate our framework, we use Dataset 2 of the CaloChallenge 2022. Through the integration of classical computation and quantum simulation, this hybrid framework paves way for utilizing large-scale quantum simulations as priors in deep generative models.

1 Introduction

The High-Luminosity Large Hadron Collider (HL-LHC), expected to be operational by the end of this decade, will offer unprecedented opportunities to measure the Higgs boson properties, explore the Standard Model in greater depth, while also searching for physics beyond the Standard Model [1] A critical component of this endeavor is the vast amount of data obtained from numerical simulations, which play a crucial role in both the design of future experiments and in the analysis of current ones. These simulations, done with Geant4 [2, 3], accurately describe the collisions at the Large Hadron Collider (LHC). But this comes at the price of being computationally intensive. These simulations describe the interactions between detectors and primary particles, but also account for the interaction with secondary particles produced as the primary particles interact with the detector material. Such is the case with calorimeters, which measure energy deposition from showers of secondary particles. Current projection for the HL-LHC run estimate millions of CPU-years per year [4]. Simulating one single event with Geant4 in an LHC experiment requires approximately 1000 CPU seconds, with the calorimeter simulation being the most resource-intensive module [5]. Through the generation of these showers, non-negligible computational resources are being employed in keeping track of these particles. Deep generative surrogates are being developed to model the particle-calorimeter interactions in the simulation pipeline, potentially reducing the overall time to simulate single events by several orders of magnitude. Examples of these are Generative Adversarial Networks [6–8], which are now an integral part of the simulation pipeline [9, 10]. Similarly VAEs [4, 11, 12], Normalizing Flows [13, 14], Transformers [15], Diffusion models [16–18] and combinations thereof [19–22], where the last reference combines a VAE with a two-partite quantum annealer (QA). The framework combining VAE with QAs has also been used in different contexts [23, 24].

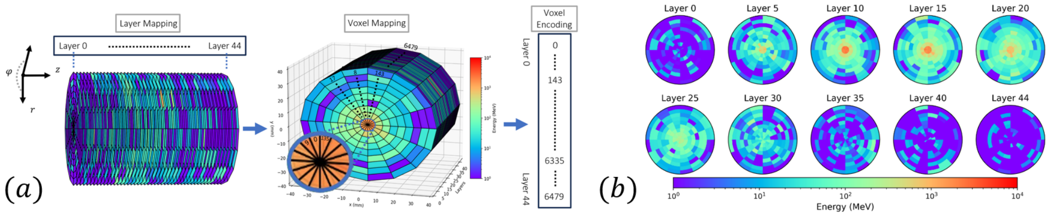

Figure 1: **(a)** Calochallenge dataset showers are voxelized using cylindrical coordinates (r, φ, z). For any given event, each voxel value corresponds to the energy (MeV) in that vicinity. Each layer has 144 voxels composed of 16 angular bins and 9 radial bins. The data set is parsed onto a 1D vector with 6480 voxels per each event. **(b)** Visualization of the voxels in an event in the dataset.

2 Methods

We illustrate our framework by using Dataset 2 of the CaloChallange-2022 [25]. This dataset consists of 100k Geant4-simulated electron showers ranging from 1 GeV to 1 TeV incident particle energy, sampled from a log-uniform distribution. The voxelized detector is in the form of a concentric cylinder with 45 layers in the axial direction of which each layer is made up of an alternating collider-absorber, active (silicon) and passive (tungsten), material. Each layer consists of 144 voxels (volumetric pixels), 9 radially and 16 in the angular direction to yield a total of 45 x 16 x 9 = 6480 voxels in one event as shown in Fig. 1. Each event has its corresponding incident particle energy as its label. We preprocess the data similar to [19], except we omit the last step where the new variable is standardized. Instead we apply a shift to the logits to preserve the sparsity of the shower in the new variables, *i.e.*, the new variable is zero whenever the voxel energy is zero.

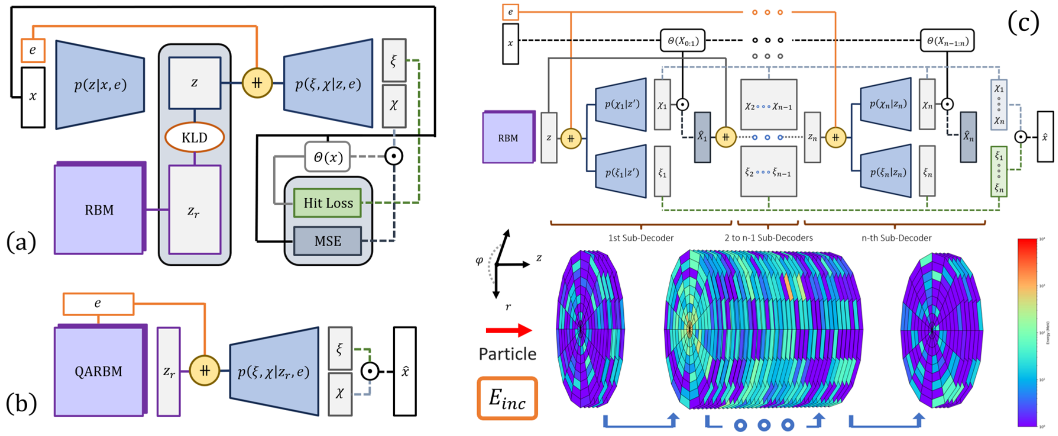

Figure 2: **(a)** Overview of Calo4pQVAE training architecture. Preprocessed voxels of a shower, x and their corresponding incident energies, e are inputted to the encoder. The encoder compresses the energy-conditioned shower into 4 partitions, of which 3 are generated from the hierarchies of the encoders and 1 is an encoded conditioning of the incident energy. The conditioned RBM is trained to learn these representations, then the concatenation of the 4 partitions is the latent vector that gets passed through the hierarchical decoder, generating a hits and activations vector to reconstruct the shower. **(b)** Once the model finishes training classically, the states of the trained RBM with an incidence energy conditioning is loaded onto D-Wave's Zephyr quantum annealer to sample a latent vector that is then passed through the hierarchical decoder to generate a shower. **(c)** The hierarchical decoder consists of 9 sub-decoders, each generating 5 layers to make up a total of 45 layers and conditions subsequent layers of the shower based on previous layers to simulate the physical propagation of particle scattering in the calorimeter through the evolution of the shower.

Our model is a variational autoencoder [26] with a 4-partite conditioned restricted Boltzmann machine [27] as the prior, as illustrated in Fig. 2 (a) . We used a hierarchical encoder composed of three sub-encoders. These hierarchy levels enforce couplings among latent units by introducing conditioning among latent nodes. In addition, these hierarchy levels introduce skip connections akin to residual networks [28]. We feed the encoded sample from each of the three sub-encoder outputs to three of the partitions in the RBM, while the fourth partition is used to condition the RBM. The RBM condition parameter is the binarized incident particle energy of the event. The prior is the 4-partite restricted Boltzmann machine, where the connections between nodes mimic the Zephyr topology of D-wave's QA [29]. The encoded sample is then fed to the hierarchical decoder, as shown in Fig. 2 (c) where n sub-decoders are allocated to generate $45/n$ layers per sub-decoder. The hierarchical decoder conditions subsequent layers of the shower based on previous layers through hierarchies of auto regressive sub-decoders to simulate the physical propagation of particle scattering in the calorimeter during the evolution of a shower. The hierarchical decoder in Calo4pQVAE consists of 9 subdecoders, each generating 5 layers, making up the entire 45-layer voxelized shower. The decoder outputs a mask vector and an activation vector, and their Hadamard product yields the generated shower. We use the Gumbel trick [30] in our framework to generate both the encoded shower as well as to generate the output mask. Our code is publicly available and can be found here [31].

3 Results

We trained our model classically for 100 epochs via the evidence lower bound (ELBO) function similar to [21], set the number of Gibbs sampling steps for the RBM to 3000 and used contrastive

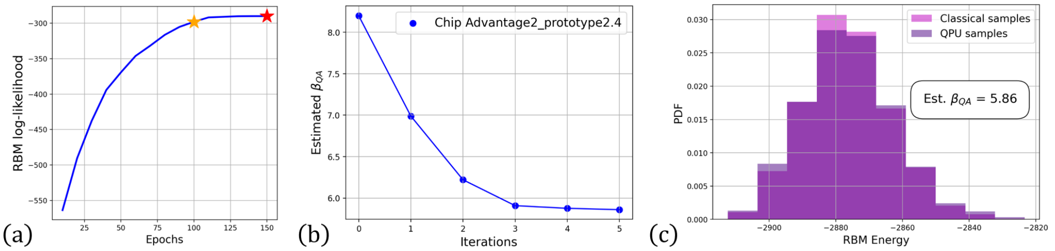

Figure 3: **(a)** Saturation of RBM log-likelihood vs epochs. Yellow star - freezing of encoder and decoder, Red star - completion of model training. **(b)** QA inverse temperature estimation *vs* iterations. **(c)** RBM energy histogram for classical and QA samples.

divergence [32]. During the first 45 epochs we linearly annealed the activation function slope used in the Gumbel trick, from 5 to 500. Afterwards, we continued the training for another 45 epochs, afterwards we froze the encoder and decoder parameters and continued training the prior up to 150 epochs in total. The model was trained using NVIDIA RTX A6000. We validate our model using D-wave's Advantage2_prototype2.3 for inference. It has been well documented how QAs can reach a freeze-out state [33], akin to glass-forming melts under a fast quench [34]. Despite this, it has been shown that the distribution in this freeze-out state can be approximated with a Boltzmann distribution [23]. We estimate the effective inverse temperature of the QA by means of a mapping with an attractive fix point at the QA's effective inverse temperature. This mapping is robust and converges faster than the method used in [21]. In Fig. 3 we show, *(a)*, RBM log-likelihood vs epochs estimated via (reverse) annealed importance sampling [35, 36], *(b)*, a set of iterations to estimate the QA effective inverse temperature and, *(c)*, the RBM energy histograms obtained from classical Monte Carlo and QA sampling.

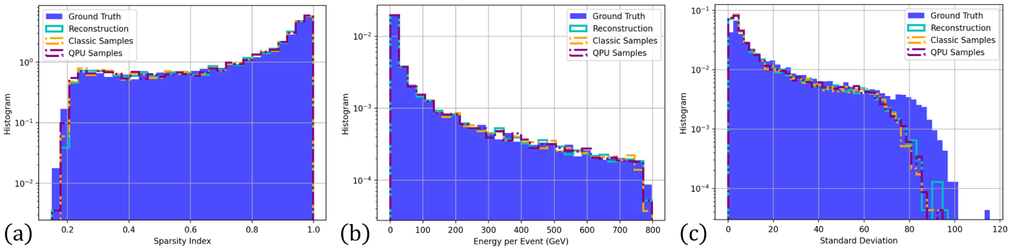

Figure 4: Normalized histograms comparing Geant4 simulated data (ground truth) and Calo4pQVAE's reconstruction, classically sampled synthetic data, and quantum annealed (Zeyphr) synthetic data for 10k events in: **(a)** sparsity index, ratio of non-hit voxels over all voxels in a shower, **(b)** energy per event, sum of all voxel energies in a shower, and **(c)** granularity, randomly shifted differences in voxel energies along angular and radial bins in a shower.

In Fig. 4 we show the histograms for sparsity index (defined as the ratio between zero-energy voxels and total number of voxels), the energy per event and the shower standard deviation of the shower angular fluctuation, for ground truth, reconstruction, classical samples and QA samples. In Fig. 5 we compare the mean energy along the radial, angular and axial axis between the ground truth and our model. In Table 1 we present the Fréchet physics distance (FPD) and the Kernel physics distance (KPD) scores between our synthetic data and the Geant4 data, using the JetNet package [37]. These values are within the limits of the models analyzed in the CaloChallenge [38]. There are additional metrics as part of the CaloChallenge that we will consider as an immediate continuation of this work.

Model	FPD ($\times 10^{-3}$)	KPD ($\times 10^{-3}$)
Calo4pQVAE	1399.48 ± 9.70	15.94 ± 1.02

Table 1: FPD and KPD values for Calo4pQVAE.

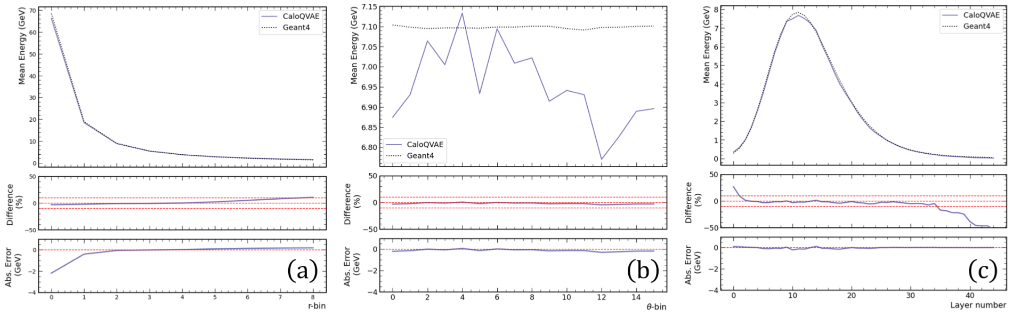

Figure 5: Solid vs dotted line plots comparing Geant4 simulated data and classical Calo4pQVAE's classically sampled synthetic data for 100k events. **(a)** Mean energy deposit per layer, **(b)** angular and **(c)** radial bin. Relative and absolute errors for each parameter are shown underneath each plot, respectively.

4 Conclusion

In this paper we presented our 4-partite quantum-assisted deep generative model for calorimeter synthetic data generation. This framework provides competitive performance for simulating particle showers at the LHC experiments while running extremely quickly on D-wave quantum annealers. The quality of the synthetic data is average compared to other approaches [16, 19]. This may be due in part to the sparse connectivity of the RBM, which mimics the QA connectivity. In addition to RBM connectivity, our framework could benefit from using attention layers similar to [19], which we leave for future work. Furthermore, Ref. [22] presents an improved model that reaches KPD and FPD values of the order of $0.9 \cdot 10^{-3}$ and $450 \cdot 10^{-3}$, making our framework competitive compared to the frameworks analyzed in the CaloChallenge [38].

The generation time using GPU is dominated by the block Gibbs sampling steps to reach equilibrium. However, the number of steps to reach equilibrium is strongly dependent on training [39]. In our framework, we used 3000 steps, which is more than typically used. Under these conditions, the generation time per event using GPU is roughly 500 times faster than Geant4. Although the annealing time per sample using QPU is $20\ \mu s$, there is technical overhead. Under the previous conditions, the generation time per event using QPU is roughly one order of magnitude faster than using GPU. A more rigorous analysis is required in this comparison, due to the nuances involved in estimating the generation time using QPU and using GPU, and since our preliminary results indicate that they differ by one order of magnitude. We leave this for future work and reiterate that our framework is significantly faster than Geant4. In conclusion, our work on Calo4pQVAE demonstrates the utility of hybrid classical and quantum frameworks in generative AI. This hybrid framework opens new opportunities for leveraging large-scale quantum simulations as priors within deep generative models for high-energy physics and potentially beyond.

Acknowledgments and Disclosure of Funding

We gratefully acknowledge Jack Raymond, Mohammad Amin, Trevor Lanting and Mark Johnson for their feedback and discussions. We gratefully acknowledge funding from the National Research Council (Canada) via Agreement AQC-002, Natural Sciences and Engineering Research Council (Canada), in particular, via Grants SAPPJ-2020-00032 and SAPPJ-2022-00020. This research was supported in part by Perimeter Institute for Theoretical Physics. Research at Perimeter Institute is supported by the Government of Canada through the Department of Innovation, Science and Economic Development and by the Province of Ontario through the Ministry of Research, Innovation and Science. The University of Virginia acknowledges support from NSF 2212550 OAC Core: Smart Surrogates for High Performance Scientific Simulations and DE-SC0023452: FAIR Surrogate Benchmarks Supporting AI and Simulation Research. JQTM acknowledges a Mitacs Elevate Postdoctoral Fellowship (IT39533) with Perimeter Institute for Theoretical Physics.

References

[1] ATLAS collaboration et al. Physics at a high-luminosity lhc with atlas. *arXiv preprint arXiv:1307.7292*, 2013.

[2] GEANT Collaboration, S Agostinelli, et al. Geant4–a simulation toolkit. *Nucl. Instrum. Meth. A*, 506(25):0, 2003.

[3] John Allison, Katsuya Amako, John Apostolakis, Pedro Arce, Makoto Asai, Tsukasa Aso, Enrico Bagli, A Bagulya, S Banerjee, GJNI Barrand, et al. Recent developments in geant4. *Nuclear instruments and methods in physics research section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 835:186–225, 2016.

[4] ATLAS collaboration et al. Deep generative models for fast photon shower simulation in ATLAS. *arXiv preprint arXiv:2210.06204*, 2022.

[5] David Rousseau. Experimental particle physics and artificial intelligence. In *Artificial Intelligence for Science: A Deep Learning Revolution*, pages 447–464. World Scientific, 2023.

[6] Luke de Oliveira, Michela Paganini, and Benjamin Nachman. Learning particle physics by example: location-aware generative adversarial networks for physics synthesis. *Computing and Software for Big Science*, 1(1):4, 2017.

[7] Michela Paganini, Luke de Oliveira, and Benjamin Nachman. Accelerating science with generative adversarial networks: an application to 3d particle showers in multilayer calorimeters. *Physical review letters*, 120(4):042003, 2018.

[8] Michela Paganini, Luke de Oliveira, and Benjamin Nachman. Calogan: Simulating 3d high energy particle showers in multilayer electromagnetic calorimeters with generative adversarial networks. *Physical Review D*, 97(1):014021, 2018.

[9] ATLAS collaboration et al. Fast simulation of the ATLAS calorimeter system with generative adversarial networks. *ATLAS PUB Note, CERN, Geneva*, 2020.

[10] Georges Aad, Brad Abbott, Dale C Abbott, A Abed Abud, Kira Abeling, Deshan Kavishka Abhayasinghe, Syed Haider Abidi, Asmaa Aboulhorma, Halina Abramowicz, Henso Abreu, et al. Atlfast3: the next generation of fast simulation in ATLAS. *Computing and software for big science*, 6(1):7, 2022.

[11] Erik Buhmann, Sascha Diefenbacher, Engin Eren, Frank Gaede, Gregor Kasieczka, Anatolii Korol, and Katja Krüger. Decoding photons: Physics in the latent space of a bib-ae generative network. 251:03003, 2021.

[12] Dalila Salamani, Anna Zaborowska, and Witold Pokorski. Metahep: Meta learning for fast shower simulation of high energy physics experiments. *Physics Letters B*, 844:138079, 2023.

[13] Claudius Krause and David Shih. Caloflow: fast and accurate generation of calorimeter showers with normalizing flows. *arXiv preprint arXiv:2106.05285*, 2021.

[14] Matthew R Buckley, Ian Pang, David Shih, and Claudius Krause. Inductive simulation of calorimeter showers with normalizing flows. *Physical Review D*, 109(3):033006, 2024.

[15] Luigi Favaro, Ayodele Ore, Sofia Palacios Schweitzer, and Tilman Plehn. Calodream - detector response emulation via attentive flow matching. *arXiv preprint arXiv:2405.09629*, 2024.

[16] Vinicius Mikuni and Benjamin Nachman. Caloscore v2: single-shot calorimeter shower simulation with diffusion models. *Journal of Instrumentation*, 19(02):P02001, 2024.

[17] Dmitrii Kobylianskii, Nathalie Soybelman, Etienne Dreyer, and Eilam Gross. Calograph: Graph-based diffusion model for fast shower generation in calorimeters with irregular geometry. *arXiv preprint arXiv:2402.11575*, 2024.

[18] Qibin Liu, Chase Shimmin, Xiulong Liu, Eli Shlizerman, Shu Li, and Shih-Chieh Hsu. Calo-vq: Vector-quantized two-stage generative model in calorimeter simulation. *arXiv preprint arXiv:2405.06605*, 2024.

[19] Oz Amram and Kevin Pedro. Denoising diffusion models with geometry adaptation for high fidelity calorimeter simulation. *Physical Review D*, 108(7):072014, 2023.

[20] Thandikire Madula and Vinicius M Mikuni. Calolatent: Score-based generative modelling in the latent space for calorimeter shower generation. 2023.

[21] Sehmimul Hoque, Hao Jia, Abhishek Abhishek, Mojde Fadaie, J Quetzalcoatl Toledo-Marín, Tiago Vale, Roger G Melko, Maximilian Swiatlowski, and Wojciech T Fedorko. Caloqvae: Simulating high-energy particle-calorimeter interactions using hybrid quantum-classical generative models. *arXiv preprint arXiv:2312.03179*, 2023.

[22] J Quetzalcoatl Toledo-Marin, Sebastian Gonzalez, Hao Jia, Ian Lu, Deniz Sogutlu, Abhishek Abhishek, Colin Gay, Eric Paquet, Roger Melko, Geoffrey C Fox, et al. Conditioned quantum-assisted deep generative surrogate for particle-calorimeter interactions. *arXiv preprint arXiv:2410.22870*, 2024.

[23] Walter Winci, Lorenzo Buffoni, Hossein Sadeghi, Amir Khoshaman, Evgeny Andriyash, and Mohammad H Amin. A path towards quantum advantage in training deep generative models with quantum annealers. *Machine Learning: Science and Technology*, 1(4):045028, 2020.

[24] Vivek Dixit, Raja Selvarajan, Muhammad A Alam, Travis S Humble, and Sabre Kais. Training restricted boltzmann machines with a d-wave quantum annealer. *Frontiers in Physics*, 9:589626, 2021.

[25] Michele Faucci Giannelli, Gregor Kasieczka, Claudius Krause, Ben Nachman, Dalila Salamani, David Shih, Anna Zaborowska. Fast calorimeter simulation challenge 2022 - dataset 1,2 and 3 [data set]. zenodo. `https://doi.org/10.5281/zenodo.8099322`, `https://doi.org/10.5281/zenodo.6366271`, `https://doi.org/10.5281/zenodo.6366324`, 2022. Online; accessed TO FILL.

[26] Andrew D King, Alberto Nocera, Marek M Rams, Jacek Dziarmaga, Roeland Wiersema, William Bernoudy, Jack Raymond, Nitin Kaushal, Niclas Heinsdorf, Richard Harris, et al. Computational supremacy in quantum simulation. *arXiv preprint arXiv:2403.00910*, 2024.

[27] Geoffrey E Hinton. A practical guide to training restricted Boltzmann machines. In *Neural Networks: Tricks of the Trade: Second Edition*, pages 599–619. Springer, 2012.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. pages 770–778, 2016.

[29] Kelly Boothby, Andrew D King, and Jack Raymond. Zephyr topology of d-wave quantum processors. *D-Wave Technical Report Series*, 2021.

[30] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

[31] QaloSim. Caloqvae. `https://github.com/QaloSim/CaloQVAE`, 2024.

[32] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

[33] Mohammad H Amin. Searching for quantum speedup in quasistatic quantum annealers. *Physical Review A*, 92(5):052323, 2015.

[34] J Quetzalcóatl Toledo-Marín, Isaac Pérez Castillo, and Gerardo G Naumis. Minimal cooling speed for glass transition in a simple solvable energy landscape model. *Physica A: Statistical Mechanics and its Applications*, 451:227–236, 2016.

[35] Ruslan Salakhutdinov. Learning and evaluating boltzmann machines. *Utml Tr*, 2:21, 2008.

[36] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Accurate and conservative estimates of mrf log-likelihood using reverse annealing. In *Artificial Intelligence and Statistics*, pages 102–110. PMLR, 2015.

[37] Raghav Kansal, Javier Duarte1, Carlos Pareja, Lint Action, Zichun Hao, mova. jet-net/jetnet: v0.2.3.post3. `https://doi.org/10.5281/zenodo.5597892`, 2023. Online; accessed July 2024.

[38] Claudius Krause, Michele Faucci Giannelli, Gregor Kasieczka, Benjamin Nachman, Dalila Salamani, David Shih, Anna Zaborowska, Oz Amram, Kerstin Borras, Matthew R Buckley, et al. Calochallenge 2022: A community challenge for fast calorimeter simulation. *arXiv preprint arXiv:2410.21611*, 2024.

[39] Aurélien Decelle, Cyril Furtlehner, and Beatriz Seoane. Equilibrium and non-equilibrium regimes in the learning of restricted boltzmann machines. *Advances in Neural Information Processing Systems*, 34:5345–5359, 2021.