Energy-based graph autoencoders for semi-visible jet tagging in the Lund representation

Annapaola de Cosa[1], Roberto Seidita[1], Florian Eble[1], Christoph Ribbe[1]
[1]Department of Physics, ETH Zurich, CH

1 Semivisible Jets

Hidden Valley models provide a possible framework for dark matter [1]. They extend the Standard Model (SM) by a dark sector. If the interaction in the dark sector is confining, bound states (dark hadrons) will be formed. A fraction of these may be stable, and thus undetectable, while others may decay back into the SM. This gives rise to semivisible jets (SVJs) [2].

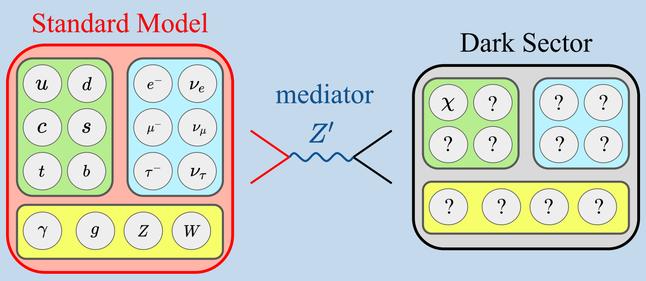

Figure 1: Hidden Valley Models.

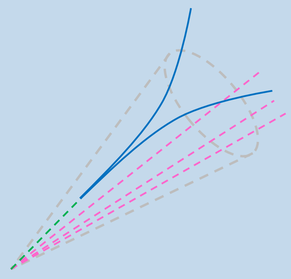

Figure 2: Semivisible Jet

2 The Lund Representation and Graph Autoencoders

Being an unordered set of particles, a jet is most naturally represented as a graph. Lund graphs specifically are chosen, since they provide a way to encode the complete clustering history of the jet [3].

An unsupervised approach is preferred for tagging these jets, as the true nature of the dark sector is unknown. Therefore, graph autoencoders are the method of choice.

Similar to a conventional autoencoder (AE), a graph autoencoder (GAE) embeds a given graph in a bottleneck dimension and subsequently reconstructs it again. The loss function is the difference between the input and output (reconstruction error).

The GAE is trained only on SM (background) jets, and is thus expected to perform worse on jets outside of the training dataset, i.e., SVJs.

When trained on SM jets, the reconstruction error can thus be used to classify between them and SVJs.

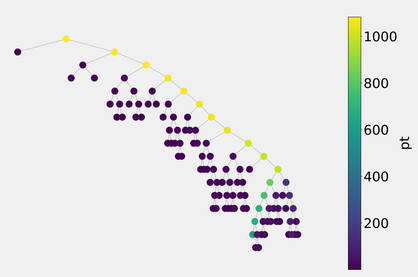

Figure 3: Lund graph of a jet with constituent pt.

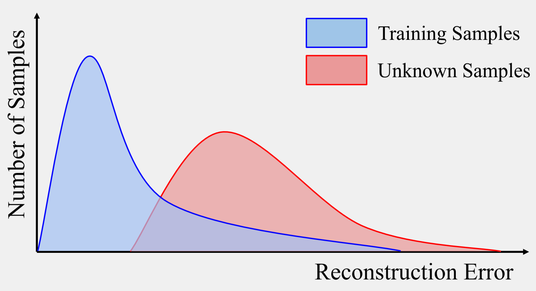

Figure 4: Loss Distribution.

3 The Problem of Outlier Reconstruction

Autoencoders are trained to have low reconstruction error on the training distribution. There is no constraint on the performance in other areas of phase space. This can lead to signal examples being reconstructed as well as background ones, limiting the ability of the AE to separate the two.

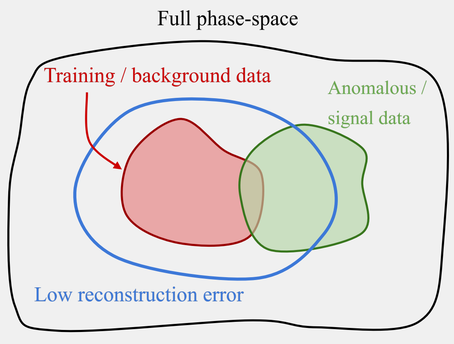

Figure 5: Outlier Reconstruction.

4 Energy-based Autoencoders

Normalized AEs (NAEs) [4] provide a mechanism to suppress outlier reconstruction. This is done by enforcing that what is learned is only the probability of the data p_{data}. This is achieved by sampling the phase space in which the AE has low reconstruction error via a Boltzmann distribution:

$$p_\theta(x) = \frac{1}{\Gamma_\theta} \exp\left(-E_\theta(x)\right),$$

where the energy E_θ is taken to be the reconstruction error of the AE, which depends on the parameters θ of the network.

The AE is then trained not to minimize the reconstruction error on the training data, but rather to match the energies on examples drawn from p_{data} and p_θ, respectively:

$$\mathbb{E}_{x\sim p_{data}}[L_\theta(x)] = \underbrace{\mathbb{E}_{x\sim p_{data}}[E_\theta(x)]}_{\text{Positive Energy}} - \underbrace{\mathbb{E}_{x'\sim p_\theta}[E_\theta(x')]}_{\text{Negative Energy}}.$$

This prescription ensures that the phase space with low reconstruction error matches the support of the training data set, suppressing outlier reconstruction in a fully unsupervised way. NAEs have been shown to be effective in separating SVJs from SM jets [5].

5 Energy-based Graph Autoencoders

Essential for this approach is the sampling of graphs from p_θ using MCMC methods. Given a graph as its feature matrix X and adjacency matrix A, the MCMC step is defined as [6]:

$$X^k = X^{k-1} - \alpha E_\theta(X^{k-1}, A^{k-1}) + \beta\omega^k$$
$$A^k = A^{k-1} - \gamma E_\theta(X^{k-1}, A^{k-1}) + \delta\eta^k,$$

where α, β, γ and δ are customizable stepsizes and ω^k and η^k are gaussian noise terms.

We show that the MCMC is able to correctly sample the space of input graphs given an example energy function, enabling the extension of the energy based paradigm to graph networks.

Figure 6: Visualization of Energy Function and MCMC.

References

[1] M. J. Strassler and K. M. Zurek. Echoes of a hidden valley at hadron colliders. *Physics Letters B*, 651(5):374–379, 2007.

[2] T. Cohen, M. Lisanti, H. K. Lou, and S. Mishra-Sharma. LHC searches for dark sector showers. *Journal of High Energy Physics*, 2017(11), nov 2017.

[3] F. A. Dreyer and H. Qu. Jet tagging in the lund plane with graph networks, 2021.

[4] S. Yoon, Y.-K. Noh, and F. Park. Autoencoding under normalization constraints. In *ICML*, pages 12087–12097. PMLR, 2021.

[5] CMS collaboration. Signal-agnostic Optimization of Normalized Autoencoders for Model Independent Searches. https://cds.cern.ch/record/2871591. 2023.

[6] M. Liu, K. Yan, B. Oztekin, and S. Ji. Graphebm: Molecular graph generation with energy-based models, 2021.