GSOC Proposal: Semi-supervised Discovery of Symmetries in CMS Calorimeter Data using Latent Flows

Contributor: [Your Name]

Mentor(s): Diptarko Choudhury, Sergei Gleyzer, Ruchi Chudasama, Samuel Campbell, Emanuele Usai, Alex Roman

Organization: ML4Sci / University of Alabama

Project: Semi-supervised Symmetry Discovery

1. Summary

This project tackles the challenge of discovering potentially hidden symmetries within complex, high-dimensional data from the CMS experiment's calorimeter stream. Standard physics symmetries, often clear in 4-vector representations, can become obscured in raw detector data or abstract feature spaces. We propose a semi-supervised deep learning approach using Variational Autoencoders (VAEs) to learn effective low-dimensional latent representations of calorimeter data. Subsequently, inspired by Oracle-Preserving Latent Flow methodologies, we will develop and train a system to identify continuous symmetry transformations (and their Lie algebra generators) within this latent space by requiring them to preserve relevant physics quantities (acting as the 'oracle'). Deliverables include a documented Python pipeline for CMS data preprocessing, VAE training, symmetry discovery, benchmarking against known symmetries or baseline methods, potentially building a physics-aware network using the discovered symmetries, comprehensive visualizations, and a final report detailing the findings.

2. Motivation and Problem Statement

Symmetries are cornerstones of the Standard Model (SM) and fundamental tools in particle physics analysis. However, representing and exploiting these symmetries directly within raw, high-dimensional detector data, such as the energy deposits in the CMS calorimeter, is challenging. Symmetries that are well-defined in terms of 4-vectors can become highly non-linear or 'hidden' when viewed through the lens of detector readouts or abstract representations learned by machine learning models.

Discovering these symmetries automatically, even when hidden, offers significant advantages. It can reveal underlying physics principles encoded in the data, lead to more robust and data-efficient ML models for tasks like particle identification or event classification by incorporating these symmetries as inductive biases, and improve model interpretability. Traditional physics analysis often relies on known symmetries, but data-driven discovery could uncover unexpected approximate symmetries or simplify analyses in complex feature spaces.

This project aims to bridge this gap by developing a machine learning framework capable of learning symmetry transformations directly from CMS calorimeter data, focusing on semi-supervised techniques that leverage the structure of the data itself and the preservation of key physics observables.

3. Goals and Deliverables

The primary goal is to develop, implement, and evaluate a semi-supervised deep learning pipeline for discovering continuous symmetries in CMS calorimeter data representations.

Deliverables:

1. **Data Handling Module:** Python code for loading, preprocessing, and preparing CMS calorimeter data samples suitable for deep learning models (addressing challenges like sparsity and high dimensionality).
2. **Representation Learning (VAE):** A trained (likely Convolutional) VAE model capable of learning a compressed and informative latent representation of the calorimeter data. Includes training infrastructure, appropriate loss functions, and model persistence.
3. **Symmetry Discovery Module (Oracle-Preserving Flow):** Implementation of the core discovery methodology:
    - Identification and implementation of suitable physics-based 'oracles' (conserved or statistically preserved quantities relevant to calorimeter data, e.g., invariant mass features, jet properties, particle type likelihoods).
    - An invertible flow model mapping the VAE latent space.
    - A generator network predicting Lie algebra parameters for transformations in the flowed space.
    - The composite loss function and stable training loop for discovering oracle-preserving transformations.
    - Analysis of the discovered generator(s) and corresponding symmetries.
4. **Benchmarking & Evaluation:** Quantitative evaluation of the discovered symmetries. This includes:
    - Verifying if known symmetries (if applicable in the chosen representation) are recovered.
    - Measuring the degree to which discovered transformations preserve the oracle quantity.
    - Comparing the performance (e.g., data efficiency, robustness) of models incorporating discovered symmetries against baseline models or potentially other symmetry-discovery techniques (as suggested by GSOC task ideas).
5. **Physics-Aware Network (Advanced/Bonus):** Using the discovered Lie algebra generator(s) to construct an equivariant/invariant neural network (potentially using EMLP or similar libraries) for a relevant downstream task (e.g., particle classification, energy regression) on the calorimeter data, demonstrating the practical utility of the discovered symmetry.
6. **Visualization Suite:** Python code generating plots for:
    - Training/validation loss curves.
    - VAE latent space visualizations (PCA/t-SNE) colored by relevant physics labels (if available).
    - VAE reconstruction examples (visualizing calorimeter energy deposits).
    - Visualizations of learned latent transformations (if feasible in low dimensions).
7. **Documentation:** Well-commented code, a README explaining setup, data requirements, and usage.

8. **Final Report:** A comprehensive report detailing the project background (CMS data, symmetries), methodology, implementation details, results (including visualizations and benchmarks), challenges, and potential future directions.

4. Proposed Solution and Methodology

The project will leverage Python, PyTorch, and potentially libraries like EMLP and JAX. The core methodology adapts the VAE + Oracle-Preserving Flow approach to the context of CMS calorimeter data:

1. **CMS Data Understanding and Preprocessing:**
    - Collaborate with mentors to select appropriate CMS open data samples or simulations representing calorimeter energy deposits.
    - Develop preprocessing steps: handling zero-suppression/sparsity, normalization, potentially converting detector geometry into suitable input formats (e.g., images, point clouds, graphs).
    - Define relevant physics quantities that can serve as 'oracles' – quantities expected to be conserved or statistically preserved under certain symmetries (e.g., related to jet kinematics, particle types).
2. **VAE for Latent Representation:**
    - Implement a suitable VAE architecture (likely CNN-based, potentially graph-based depending on data format) to learn a lower-dimensional latent space `z` from the high-dimensional calorimeter data `x`.
    - Train the VAE using reconstruction loss and KL divergence regularization, potentially employing techniques like KL annealing for stability. The goal is a latent space that captures salient physics features.
3. **Supervised Learning (Optional Baseline):**
    - If known transformations exist for the chosen data representation (e.g., rotations, translations in detector coordinates), implement a supervised MLP to learn this transformation in the latent space `z` as a baseline or sanity check.
4. **Unsupervised Symmetry Discovery:**
    - **Oracle Implementation:** Implement functions to compute the chosen physics oracle quantity `O(x)` from the input data `x`.
    - **Flow & Generator:** Implement the invertible flow `f: z -> w` and the generator network `G: w -> L` as described previously (using affine coupling, stabilized generator network).
    - **Joint Training:** Train the flow `f` and generator `G` by minimizing the composite loss:
        - `Loss = beta_oracle * ||O(x) - O(Dec(f^{-1}(exp(\epsilon L)f(Enc(x)))))||^2` (Oracle Preservation)
        - `+ beta_gen * Regularization(L)` (Generator Stability/Non-triviality, e.g., `mean((norm(L)-1)^2)`)
    - Utilize techniques like gradient clipping and careful learning rate selection to ensure stable training on the complex CMS data representations.
    - Analyze the learned generator parameters `a` and scale to understand the discovered transformation(s).

5. **Benchmarking:**
    - Quantify the oracle preservation error for the discovered transformation.
    - If known symmetries were expected, compare the learned generator `L` to the theoretical one.
    - Compare the performance of downstream models (like the oracle itself, or the bonus equivariant network) trained with and without leveraging the discovered symmetry, focusing on metrics like accuracy, robustness to transformations, and potentially data efficiency.
6. **Bonus: Physics-Aware Network:**
    - If a meaningful generator `L` is discovered:
    - Use `L` to define a custom group in EMLP.
    - Construct and train an EMLP model using this group for a relevant CMS task (e.g., jet tagging, particle ID) using the VAE latent codes as input. Compare its performance to a standard MLP baseline.
7. **Visualization and Evaluation:**
    - Implement functions to visualize calorimeter data (e.g., energy deposit heatmaps) and their VAE reconstructions.
    - Visualize latent spaces (PCA/t-SNE) colored by physics labels (e.g., particle type, jet energy bins).
    - Plot loss curves and benchmark results.

5. Proposed Timeline (Approx. 12-14 Weeks for 175h)

- **Community Bonding Period (Weeks 1-3):**
    - Setup computing environment (CERN resources if applicable).
    - Detailed study of CMS calorimeter data structure and relevant physics symmetries.
    - Deepen understanding of VAEs, Flows, Lie Algebras, and key papers.
    - Refine project plan, specific oracle choices, and data samples with mentors.
- **Week 4-5: Data Handling and VAE Implementation:**
    - Implement data loading and preprocessing for selected CMS data.
    - Implement and begin training the VAE architecture.
    - Develop initial visualization for calorimeter data and reconstructions.
- **Week 6-7: VAE Tuning and Latent Space Analysis:**
    - Tune VAE hyperparameters for meaningful latent representations.
    - Implement and analyze latent space visualizations (PCA/t-SNE).
    - Implement and train the Oracle classifier on the chosen physics quantity.
- **Week 8-10: Unsupervised Discovery Implementation & Training:**
    - Implement the Flow, Generator, and joint training loop.
    - Begin training the unsupervised discovery module.
    - Iteratively debug and stabilize the training process (tune LRs, loss weights, clipping, etc.).

- **Week 11-12: Analysis and Benchmarking:**
  - Analyze the discovered generator(s) `L`.
  - Implement and run benchmarking tests (oracle preservation, comparison to known symmetries if applicable).
  - Implement visualization for the discovered transformation (if feasible).
- **Week 13: Bonus Task / Refinement:**
  - (If time permits) Implement and train the EMLP-based physics-aware network. Evaluate its performance.
  - Refine code, documentation, and visualizations based on results.
- **Week 14: Final Report and Submission:**
  - Complete the final project report.
  - Finalize code documentation and prepare submission.

(Timeline adjustable for 350h project - potentially exploring more complex flows, multiple oracles, more thorough benchmarking, or deeper physics analysis)

6. About Me

[Provide a brief background about yourself: Your university, major, relevant coursework (ML, math, physics, particle physics?), programming experience (Python, PyTorch, C++?), previous projects (especially any related to HEP or ML), and your motivation for this specific GSOC project focusing on fundamental symmetries in physics data.]

7. References

- Roman, A., Forestano, R. T., Matchev, K. T., Matcheva, K., & Unlu, E. (2023). Oracle-Preserving Latent Flows. *arXiv:2302.00806*. (Core Method Inspiration)
- Forestano, R. T., Matchev, K. T., Matcheva, K., Roman, A., Unlu, E., & Verner, S. (2023). Deep Learning Symmetries and Their Lie Groups, Algebras, and Subalgebras from First Principles. *arXiv:2301.05638*. (Related Method)
- Finzi, M., Welling, M., & Wilson, A. G. (2021). A Practical Method for Constructing Equivariant Multilayer Perceptrons for Arbitrary Matrix Groups. *arXiv:2104.09459*. (EMLP Library)
- Kingma, D. P., & Welling, M. (2013). Auto-Encoding Variational Bayes. *arXiv:1312.6114*. (VAE Background)
- Liu, Z., & Tegmark, M. (2021). Machine-learning hidden symmetries. *arXiv:2109.09721*. (Related Concept)
- Yang, J., Walters, R., Dehmamy, N., & Yu, R. (2023). Generative Adversarial Symmetry Discovery. *arXiv:2302.00236*. (Alternative GAN-based approach)
- [Optionally add 1-2 relevant papers on ML for CMS calorimeter data or general HEP ML]