Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE

Learning a Representation of New Physics Models

Tore von Schwartz, Gregor Kasieczka, David Shih

07.11.23

ML4Jets 2023

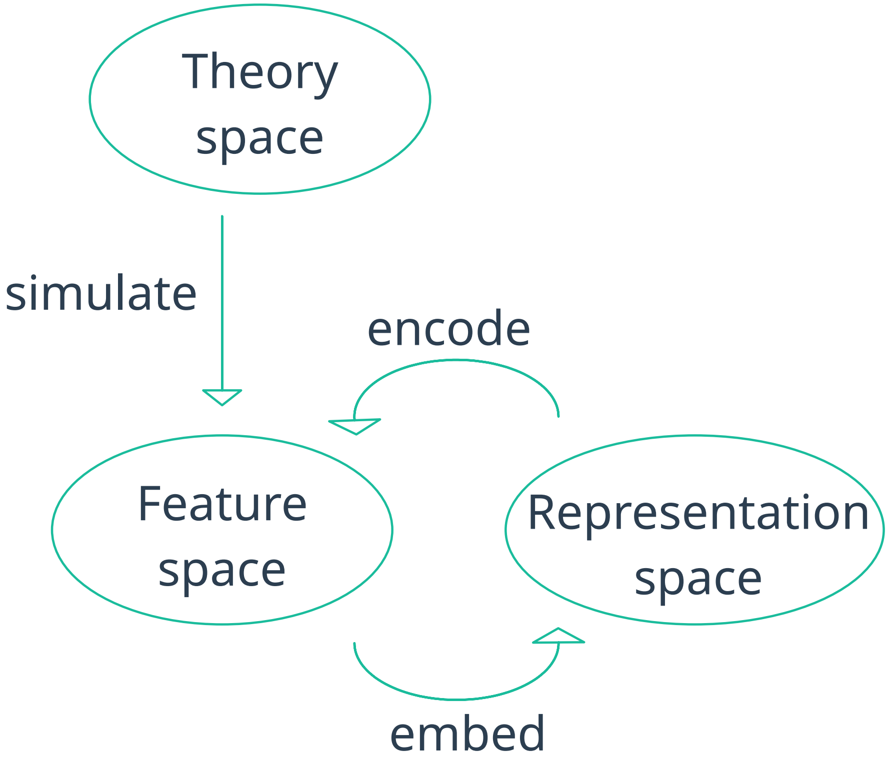
tore.von.schwartz@desy.de

Contents

- **Concept of Universal New Physics Latent Space**
- **Construction of the Embedding**
  - Autoencoder Training
  - Neural Embedding
  - Contrastive Learning
- **Application on larger selection of BSM processes**
- **Conclusion**

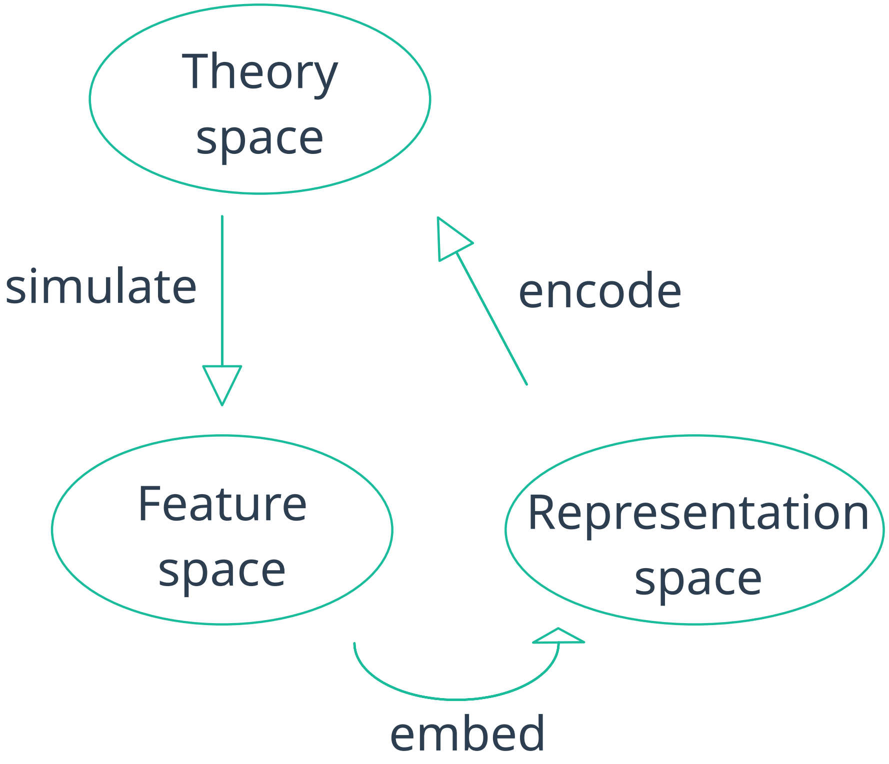
Dimensionality Reduction for Physical Data

- *„Neural Embedding: Learning the Embedding of the Manifold of Phyics Data“*, S. E. Park et al.
- Embedding in lower dimensional latent space while conserving energy mover’s distance between events



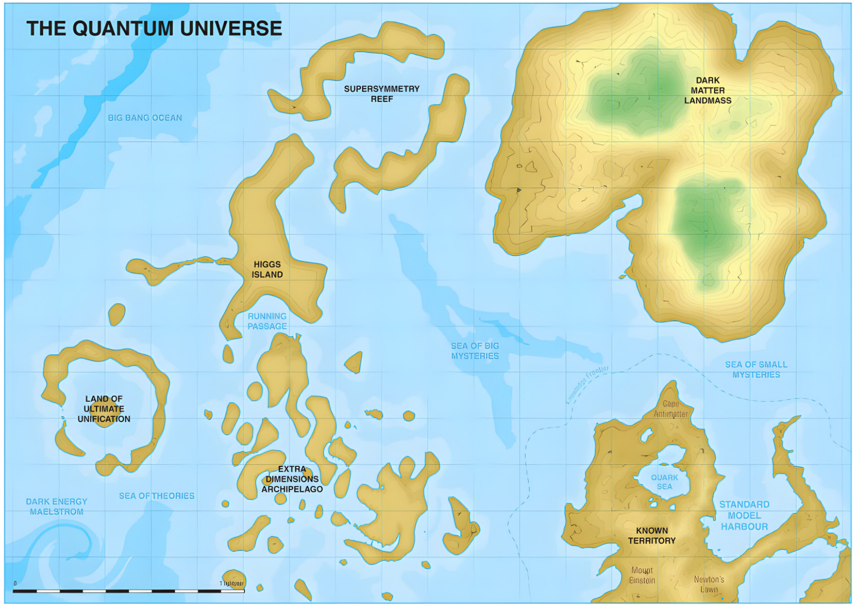
Dimensionality Reduction for Physical Data

- **Encode information about underlying theory instead of the actual events**
- **Learn embedding based on phenomenological similarities**



Universal New Physics Latent Space

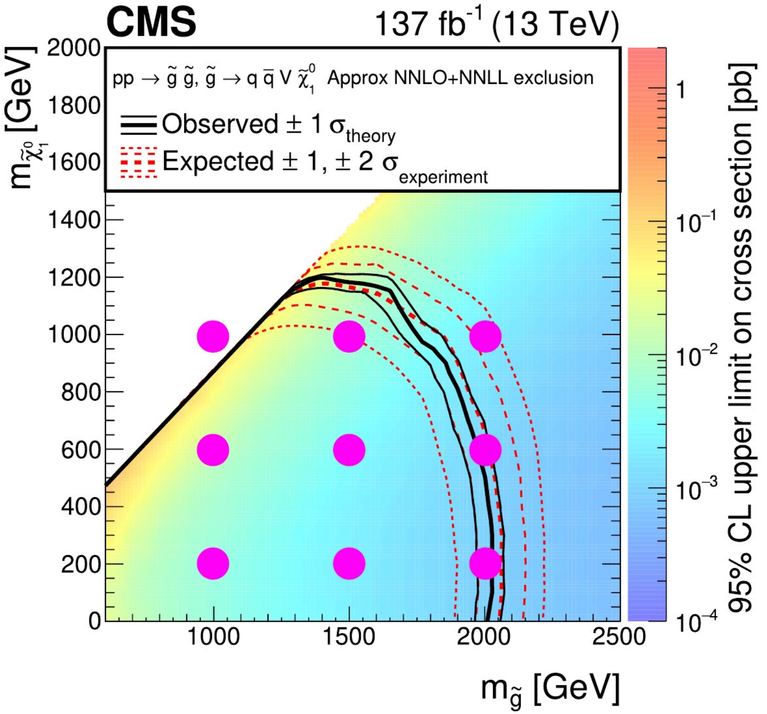
- New physics data available from simulations
- Embed data from different theories in same latent space
- Investigate phenomenological similarities in low-dimensional space

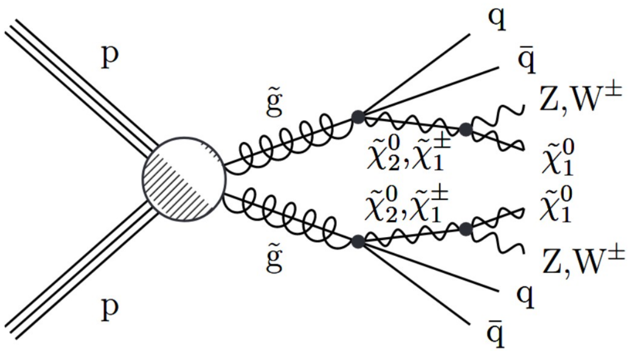


https://cds.cern.ch/record/1601971/files/ILCTDR-OUTREACH.pdf

Data Set

- **Simulation of SUSY events for different mass parameters**
- **Events are gluino decays in proton-proton collisions at 13 TeV**
- **Use kinematic features of leading four jets and missing transverse energy**

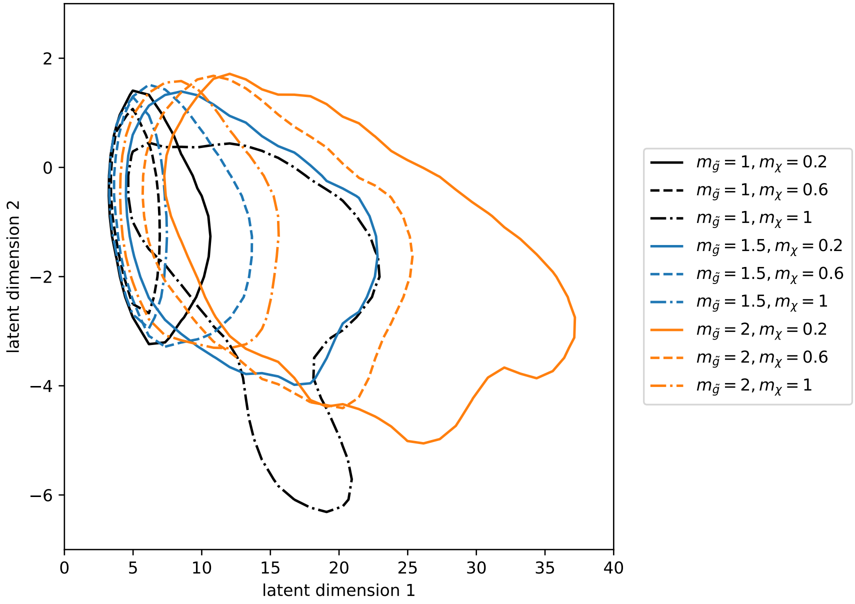




arXiv:1908.04722

First Results

- Embed data with autoencoder trained on mean squared error loss
- No clustering but slight shift in resulting structure
- Overlapping feature distributions → overlapping distributions in latent space

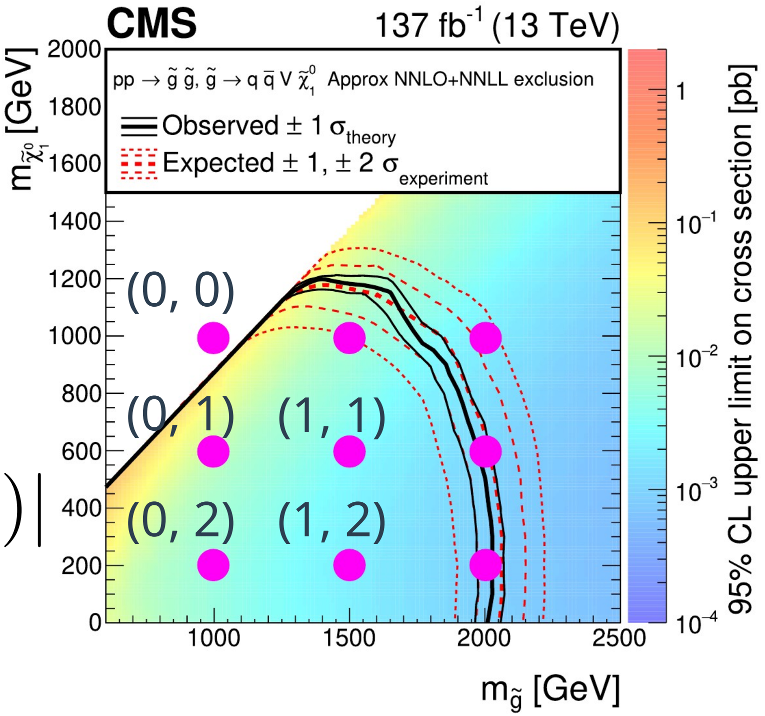


Neural Embedding

- **Additional loss term to ensure clustering**
- **Conserve metric between pair of events in latent space (arXiv:2208.05484)**

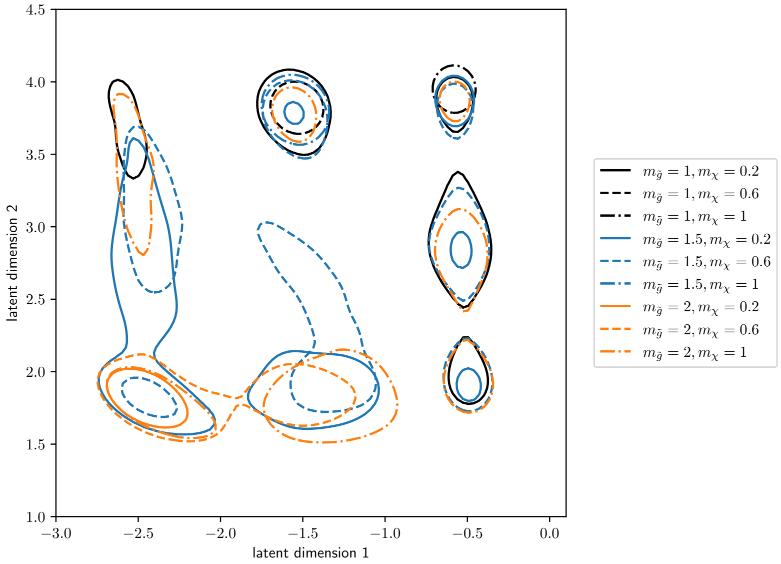
$$\mathcal{L}_{NE} \sim \left| d_{\mathcal{Y}}\left(\phi(u_i), \phi(v_i)\right) - d_{\mathcal{X}}(u_i, v_i) \right|$$

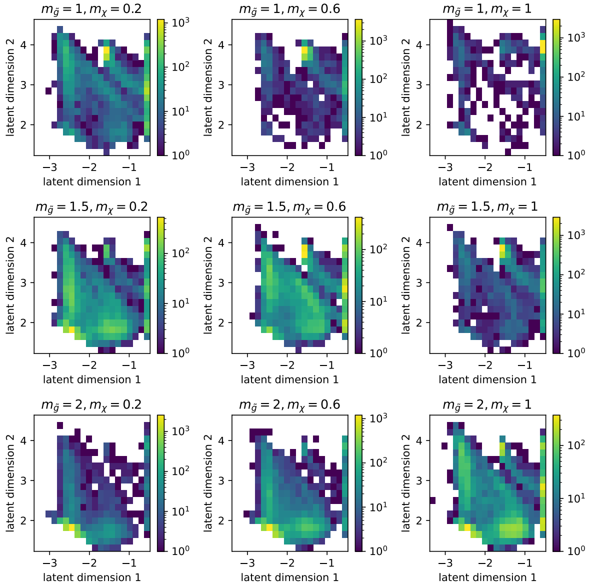
- **Here: conserve distance between data set labels**



Results Neural Embedding

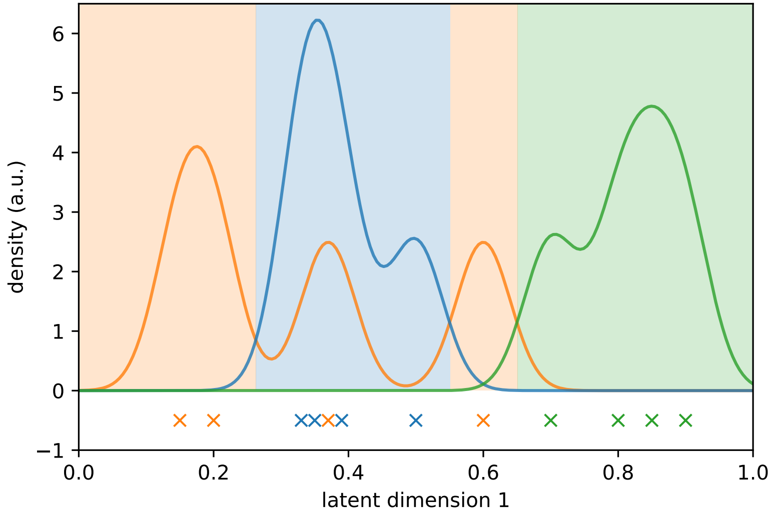
- Clustering not perfect but overall structure visible
- Arrangement of clusters similar to mass space
- Achieved embedding accuracy of ~43%





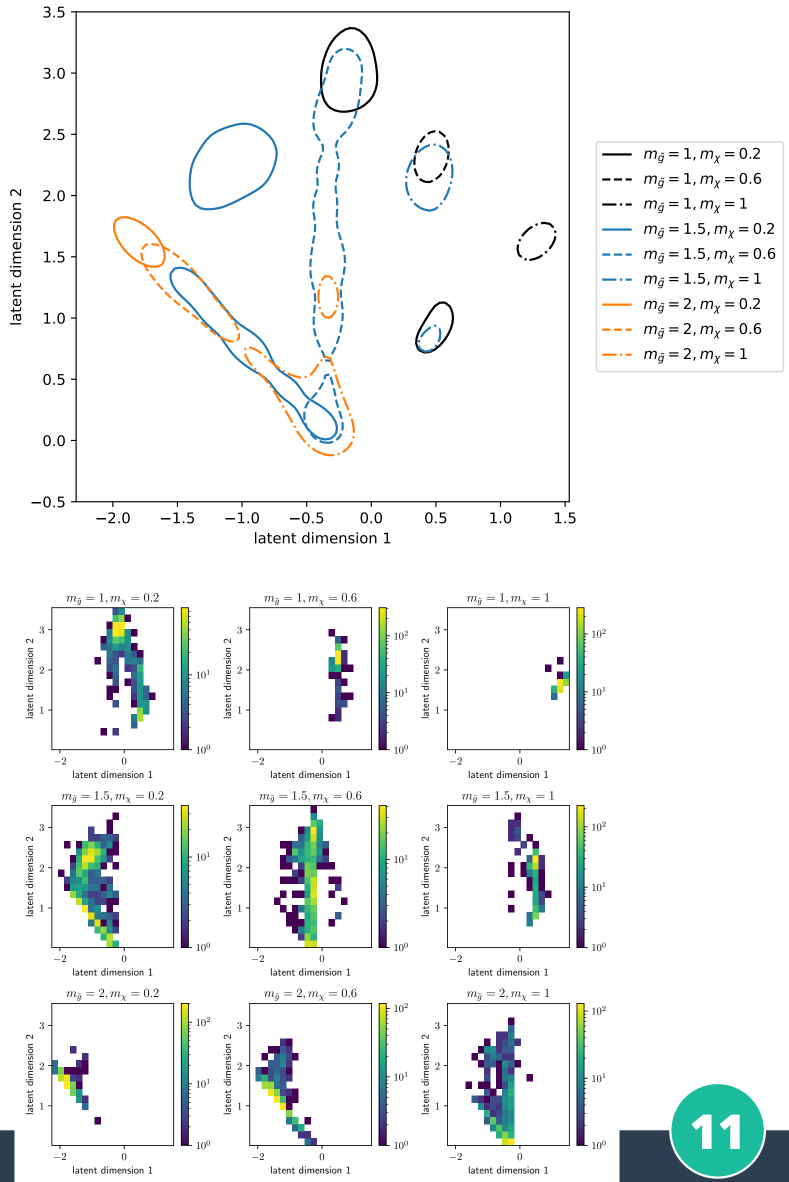
Performance Measurement

- Comparison with correct output not possible
- Assignment of latent space regions to specific classes based on kernel density estimation
- Allows calculation of "accuracy"



Training on Sets of Events

- Take sets of events from same theory as input
- Increases probability of the model to see distinguishable events
- Leads better structured latent space and accuracy of ~70%



Contrastive Learning

- Explicit and comparable data set labels not always given

  $\rightarrow$ replace NE loss term with contrastive loss term

  (Dimensionality reduction by learning an invariant mapping, R. Hadsell et al.)

$$\mathcal{L}_i = (1 - Y)\mathcal{L}_S(D_i) + Y\mathcal{L}_D(D_i)$$

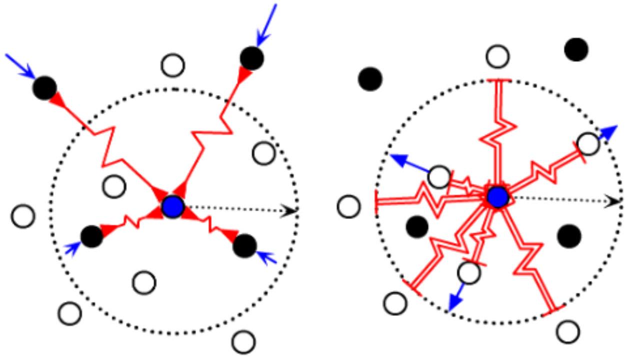
- Goal: cluster similar points and separate dissimilar points without knowing exact arrangement of data sets

Contrastive Learning

- **Exact choice of loss function**

$$\mathcal{L}_i = (1 - Y)\frac{1}{2}D_i^2 + Y\frac{1}{2}\max(0, m - D_i)^2$$

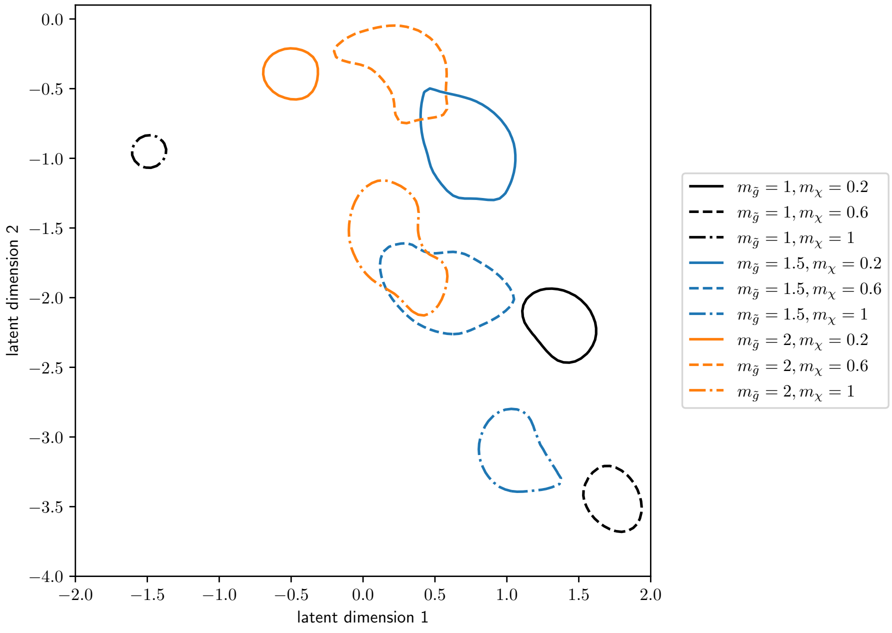
- **New margin parameter needed to deal with unbounded latent spaces**
- **Events from different theories with latent distance larger than margin parameter not longer repelled**



Dimensionality reduction by learning an invariant mapping, R. Hadsell et al.

Results Contrastive Learning

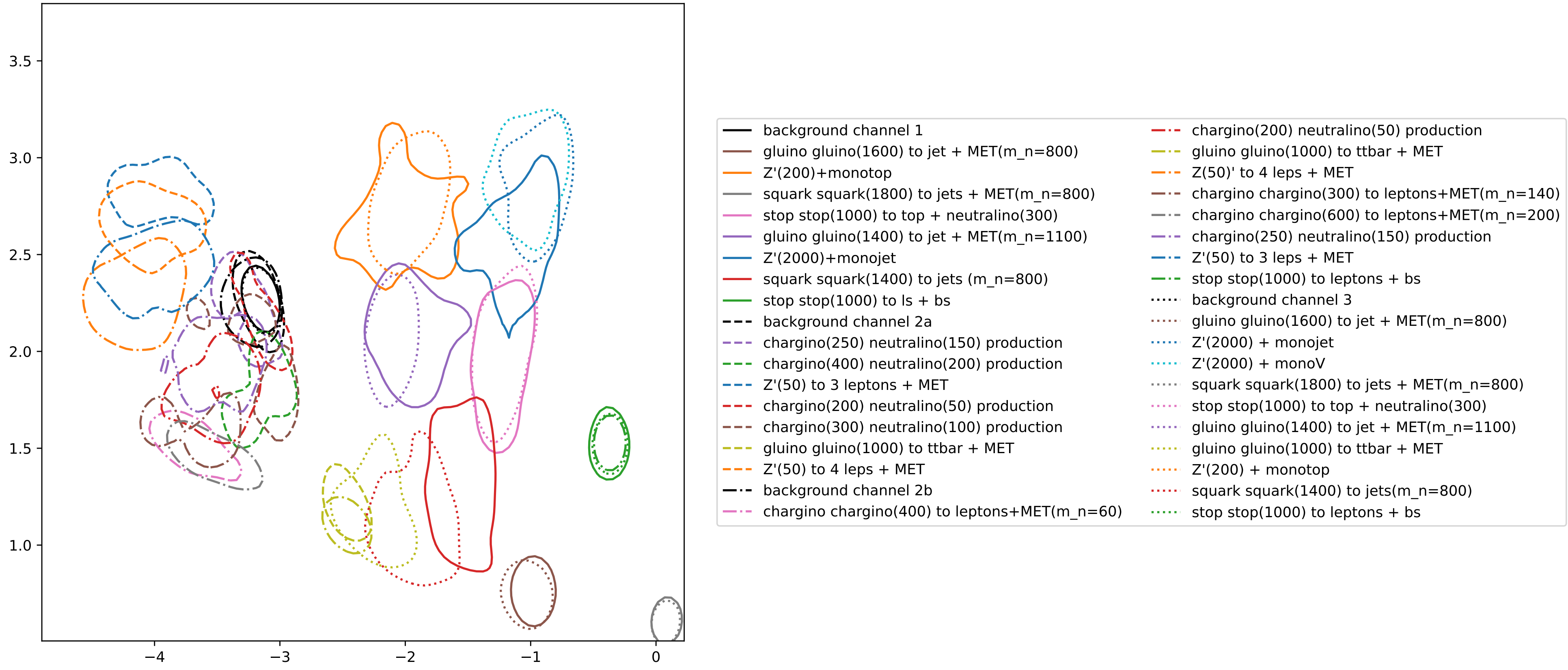
- **Clearly visible clustering**
- **Arrangement of clusters stable over multiple trainings**
  → based on physical properties
- **Unphysical data set well distinguishable**
  → almost perfectly separated



Dark Machines Data Set

- **Application of this method to Dark Machines data set including a larger variety of processes (arXiv:2105.14027)**
- **Data set divided in three channels with different focus and triggers**
  - Channel 1: hadronic activity with high missing energy
  - Channel 2a/b: leptonic events with either 2 or 3 leptons required
  - Channel 3: widest event coverage through loosest trigger
- **Background data set with SM events for every channel**

Results Dark Machines Data Set



Conclusion

- Learning an embedding based on phenomenological similarities of different theories is possible
- Training on sets as input helps to abstract from individual events
- Setup applicable to various data sets even without explicit relation in theory space