Generic representations of jets at detector-level with self-supervised learning

Kyle Cranmer, Etienne Dreyer, Eilam Gross,
Nilotpal Kakati, Dmitrii Kobylianskii, Garrett Merz,
Nathalie Soybelman, Patrick Rieck

European AI for Fundamental Physics Conference
Amsterdam, April 30th 2024

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON
NYU
WEIZMANN INSTITUTE OF SCIENCE

Self-Supervised Learning in High Energy Physics

- Supervised learning: using labeled data to find a hidden representation $h(x_{jet})$, tailored to a specific task
- Alternative: leverage unlabeled data to find a representation $h(x_{jet})$ useful for multiple tasks

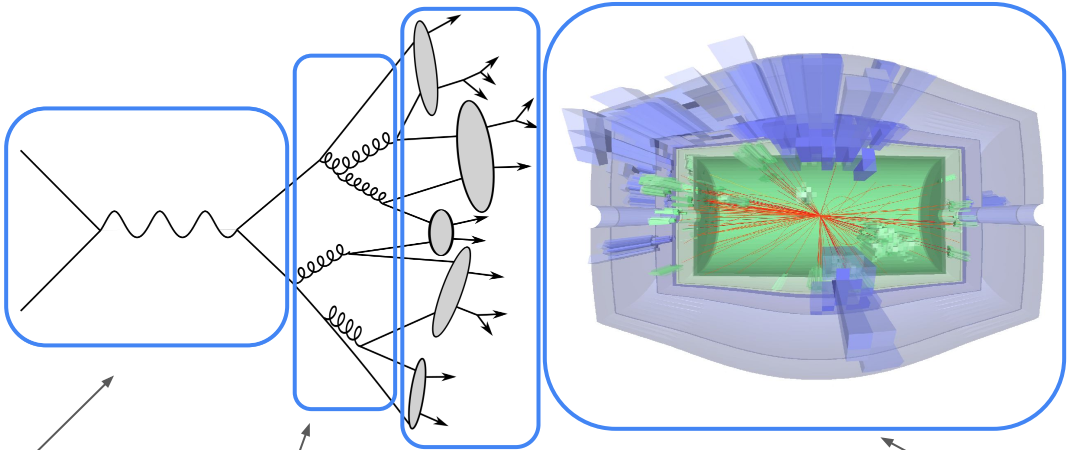
  ⇒ self-supervised learning: identify the important parts of the data, i.e. lossy compression
- One approach: pick pairs of jets incorporating the same physics of interest and require their representations to be close by

  ⇒ How to motivate notions of "sameness" ?

Related work: Symmetries, Safety, and Self-Supervision, Dillon et al. (2022),
Semi-visible jets, energy-based models, and self-supervision, Favaro et al. (2023) - talk on Thursday
Re-Simulation-based Self-Supervised Learning for Pre-Training Foundation Models, Harris et al. (2024)

Markov Process and Self-Supervised Learning

Simulation chain, Markov process:

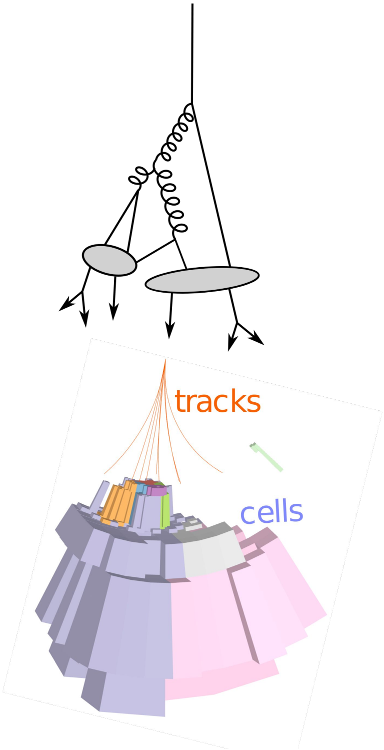


$$p(x_{\text{detector}}) = \int \mathrm{d}\mathbf{z}\; p_{\text{hard scattering}}(z_{\text{hard scattering}})\; p_{\text{parton shower}}(z_{\text{partons}}|z_{\text{hard scattering}})\; p_{\text{hadronisation}}(z_{\text{stable particles}}|z_{\text{partons}})\; p_{\text{detector}}(x_{\text{detector}}|z_{\text{stable particles}})$$

⇒ Various natural definitions of sameness of jets, set by a choice of step in the simulation chain

Different notions of Sameness

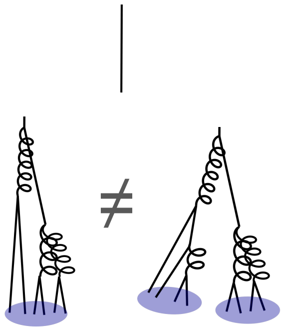
- Creation of pairs of “same” jets by running the simulation chain twice beyond a certain step



Different notions of Sameness

- Creation of pairs of “same” jets by running the simulation chain twice beyond a certain step

- One approach: rerun the parton shower
  ⇒ simplistic choice, e.g. risk of declaring 2 jets from a hard splitting as 1 jet

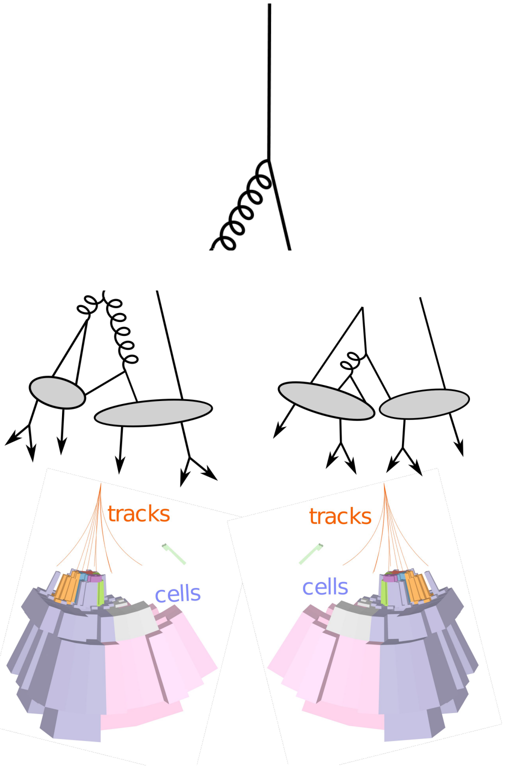


Different notions of Sameness

- Creation of pairs of “same” jets by running the simulation chain twice beyond a certain step

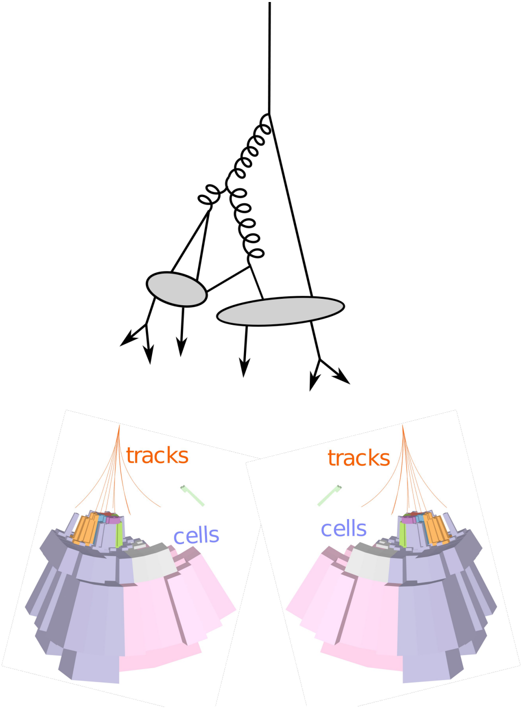
- One approach: rerun the parton shower
$\Rightarrow$ simplistic choice, e.g. risk of declaring
2 jets from a hard splitting as 1 jet

- Go deeper: rerun the simulation chain after some parton splittings



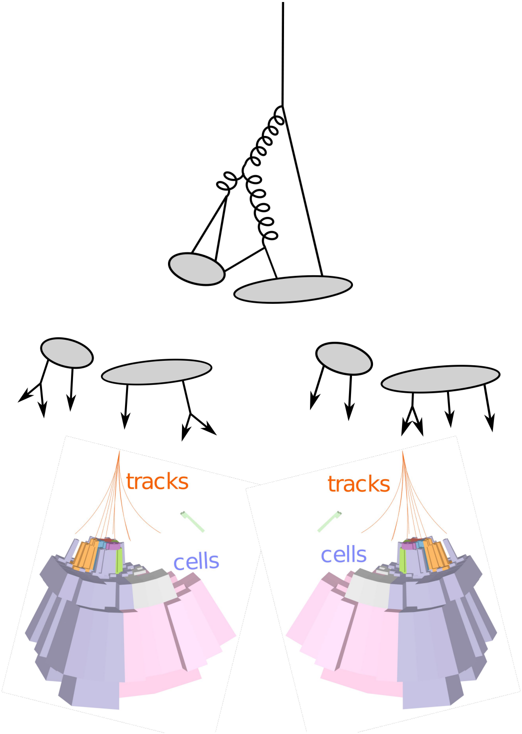
Different notions of Sameness

- Creation of pairs of “same” jets by running the simulation chain twice beyond a certain step
- One approach: rerun the parton shower
  ⇒ simplistic choice, e.g. risk of declaring 2 jets from a hard splitting as 1 jet
- Go deeper: rerun the simulation chain after some parton splittings
- Don’t go too deep: using the same particle-level jet twice gives the same tracks ⇒ collapse of the representation

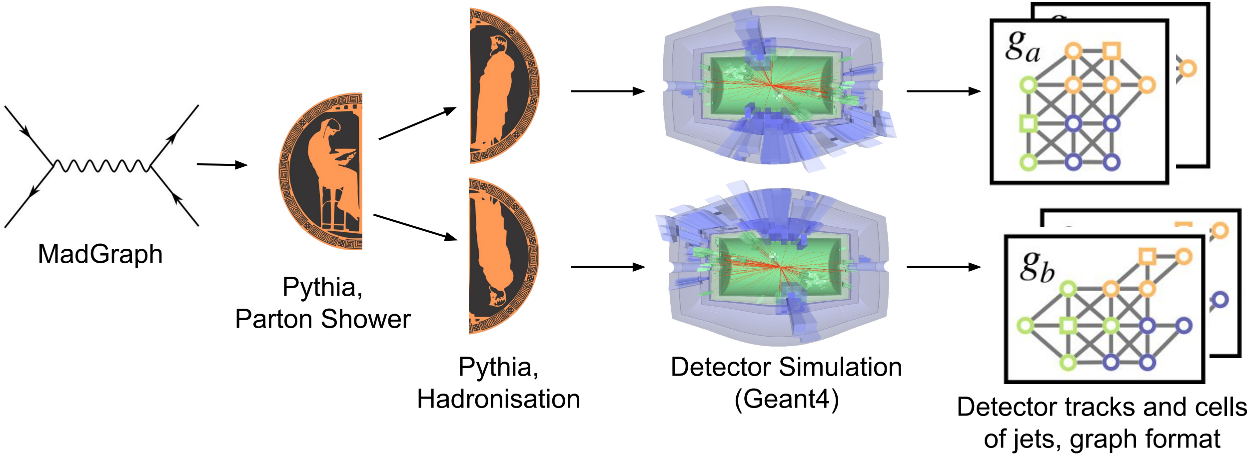


Different notions of Sameness

- Creation of pairs of “same” jets by running the simulation chain twice beyond a certain step
- One approach: rerun the parton shower
  ⇒ simplistic choice, e.g. risk of declaring 2 jets from a hard splitting as 1 jet
- Go deeper: rerun the simulation chain after some parton splittings
- Don’t go too deep: using the same particle-level jet twice gives the same tracks ⇒ collapse of the representation
- Approach in the following: frozen parton shower, only run hadronisation and detector simulation twice



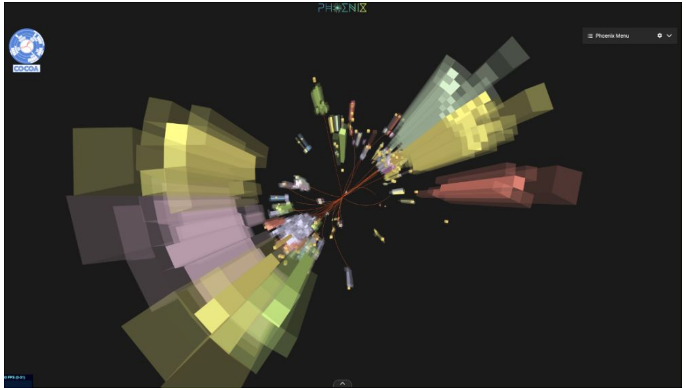
Event Simulation Chain

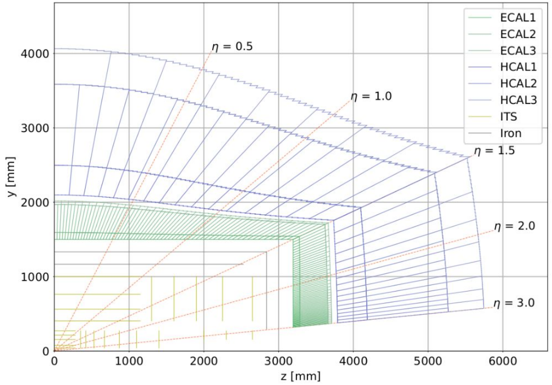


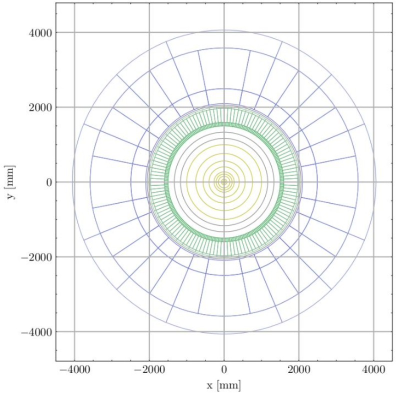
- Generation of jet events
  - Hard scattering: di-quark and di-gluon final states
  - Jet p_T approx. 100 GeV
  - Training statistics approx. 10^5 events
- Extract jets: anti-k_t algorithm, R = 0.4

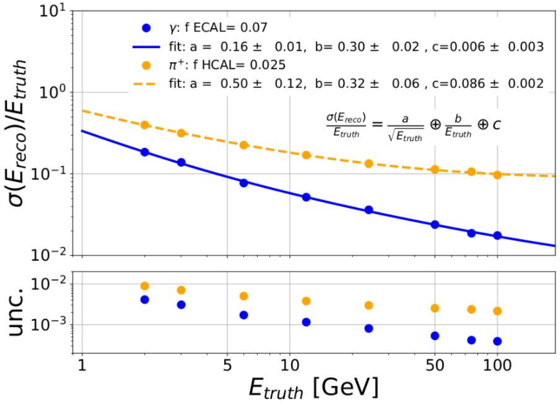
Detector Simulation

- Complicated experimental signatures of jets
  ⇒ benefit from a detailed detector simulation:
  [Cocoa](), using Geant4
- Charged particle tracker + electromagnetic and hadronic calorimeters
- Single particle calorimeter responses tuned to the ATLAS detector performance

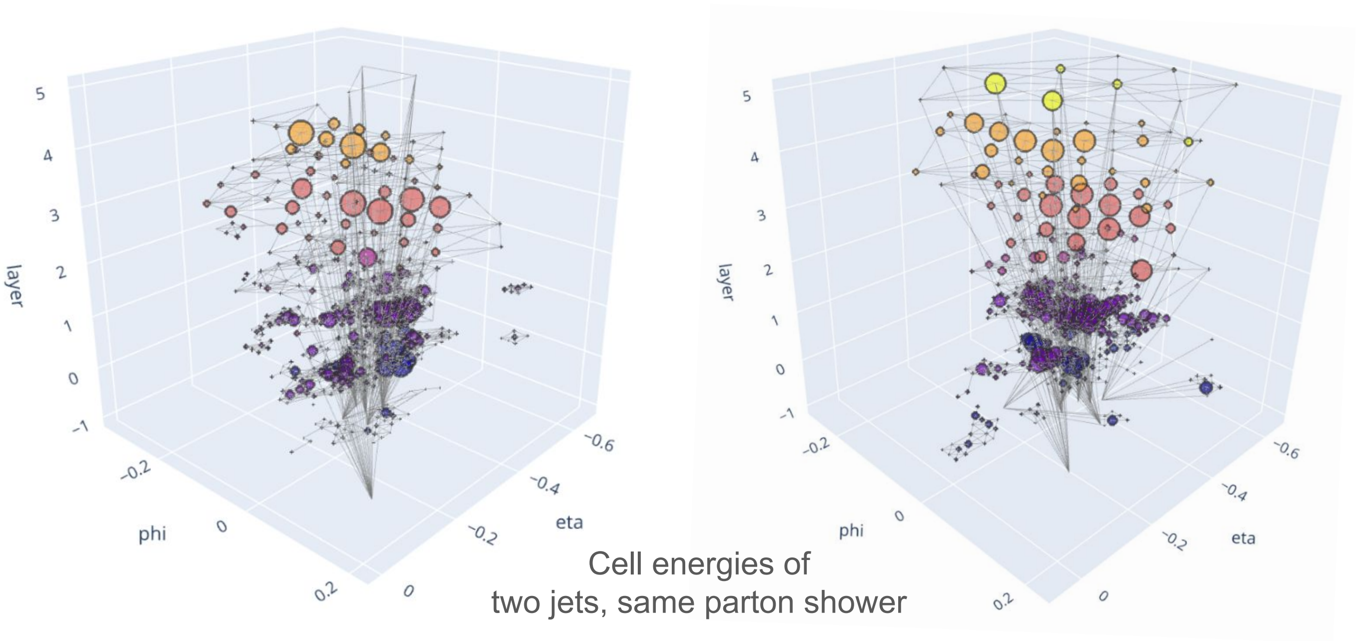








Jet tracks and cells as graphs

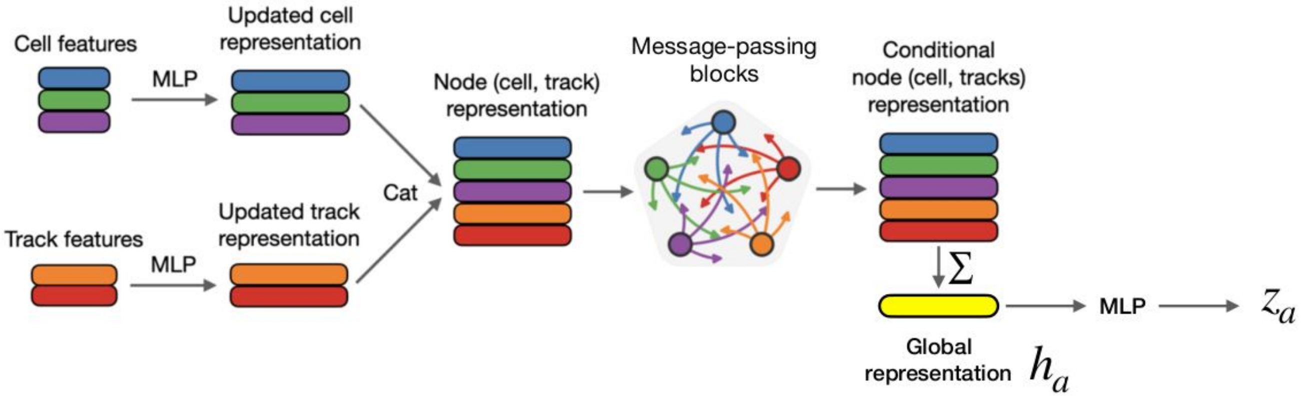


Cell energies of
two jets, same parton shower

Large variety in jet pairs due to randomness in hadronisation and detector response
⇒ non-trivial learning task

Learning Strategy

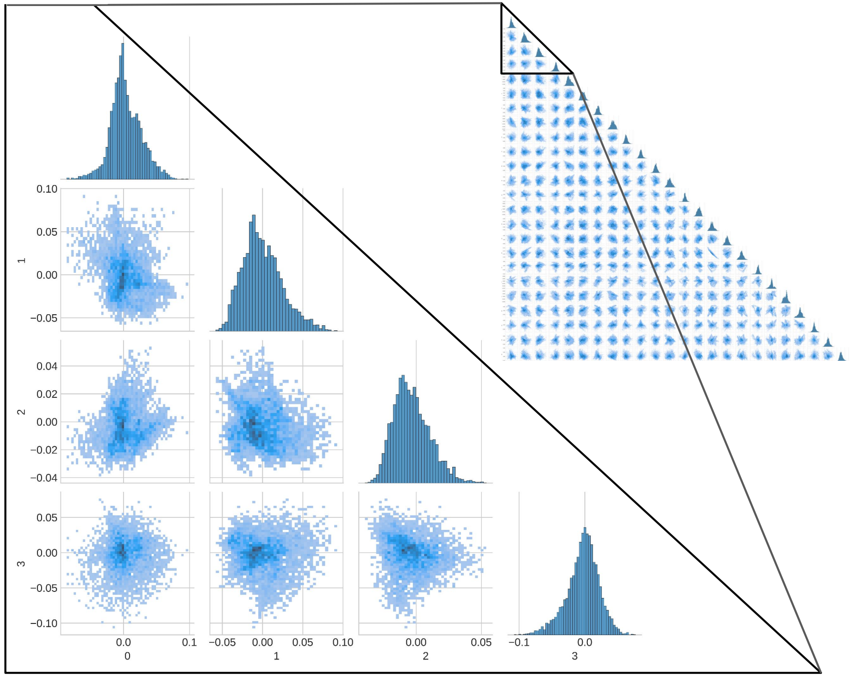
- SSL backbone: Graph neural network



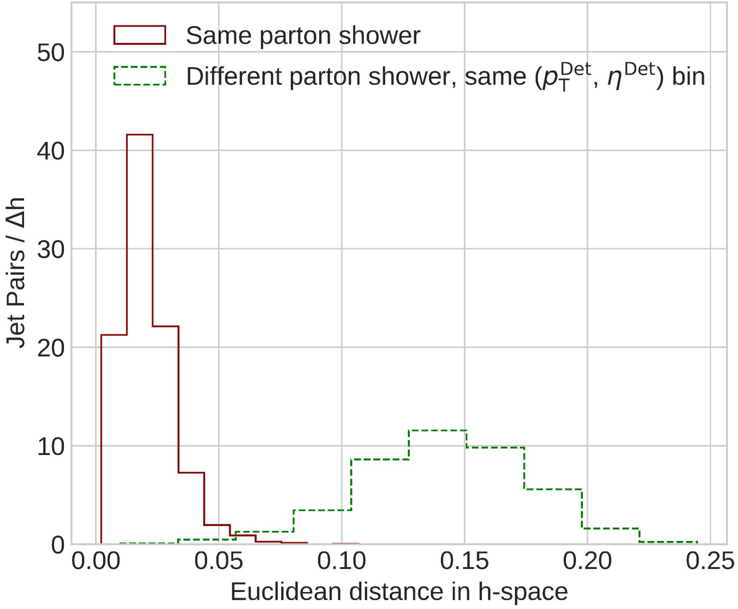
- Loss function: SimCLR

$$L(z_a, z_b) = -\log \frac{\exp(\hat{z}_a \cdot \hat{z}_b / \tau)}{\sum_{i \neq a}^{2N} \exp(\hat{z}_a \cdot \hat{z}_i / \tau)} \quad \text{where} \quad \hat{z}_a := z_a / |z_a| \implies \hat{z}_a \cdot \hat{z}_b = \cos(\theta_{ab})$$

Learned Representations

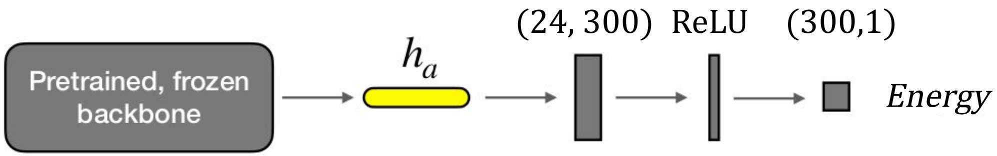


Reasonable distribution of jets in representation space

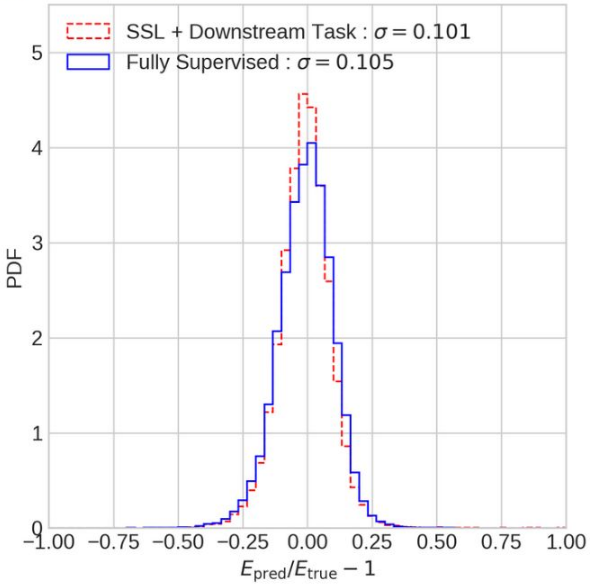


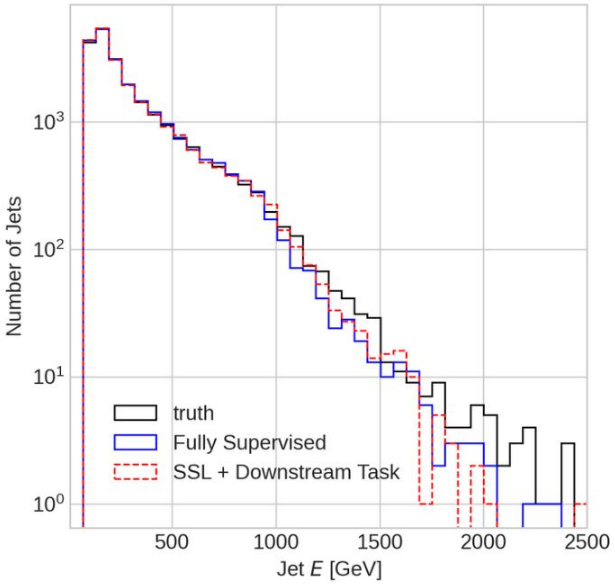
Learned to distinguish jets by their underlying parton showers!

Energy Regression

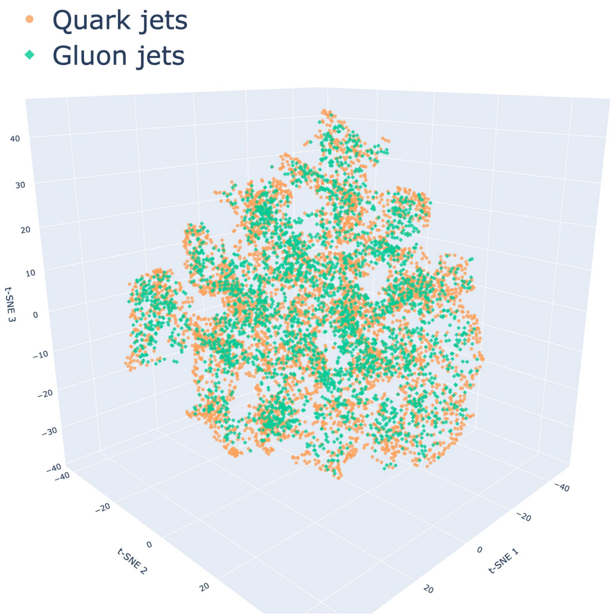
- Downstream task: 

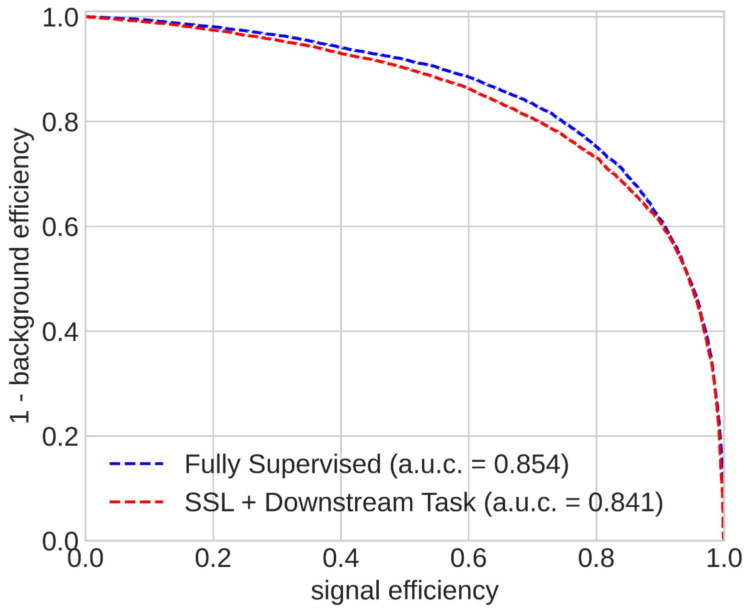
- Comparing with a fully supervised training result, same network





Quark / Gluon Tagging





- Clustering in representation space,
  SSL + kNN classifier: 73 % accuracy
- Fully supervised classifier: 78 % accuracy

Frozen SSL backbone + prediction head, compared with fully supervised classifier

Conclusion

- Built a foundation model of jets using self-supervised, contrastive learning
  - Various ways to define sameness of jets, here: frozen parton shower
- Results translate to LHC physics
  - Realistic detector simulation
  - Graph neural network