Color Singlet clustering

GNN based Reconstruction at FCC-ee

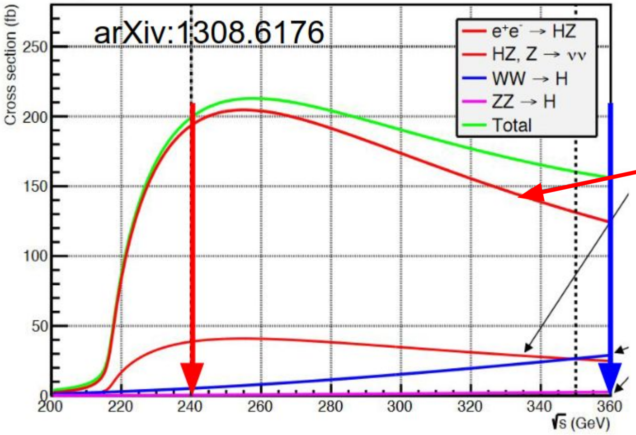
Thibault Gergaud (ENS - CERN)

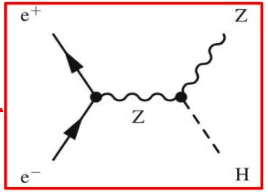
Dolores Garcia (CERN)

Michele Selvaggi (CERN)

FCC-ee

- Fcc-ee Higgs factory : produce 1.45M Higgs (HZ)





At FCC :

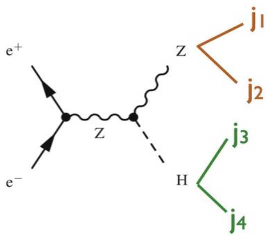
- Clean environment
- Relative small backgrounds, large S/B

ACKS : Michele Selvaggi, Dolores Garcia
https://indico.cern.ch/event/1307378/contributions/5721015/attachments/2791532/4869200/Higgs_Top_Performance_FCCWorkshop_Jan2024.pdf

HZ decay mode

Z decay m(Z) = 91Gev :
- Z(ll) ~ 10%
- Z(vv) ~ 20%
- Z(jj) ~ 70%

H decay m(H) = 125Gev :
- H(bb) ~ 58 %
- H(gg) ~ 8%
- H(tautau) ~ 6%
- H(cc) ~ 2%
- H(ss) ~ 0.02%



- Z(vv/ll) final states:
  - “easy”: 2 jets
    - from the Higgs decay
- Z(jj):
  - “hard”: 4 jets
    - can originate from H or Z

Fully hadronic, HZ → jjjj ~ 51%:
- Hardest
- Largest BR

ACKS : Michele Selvaggi, Dolores Garcia
https://indico.cern.ch/event/1307378/contributions/5724022/attachments/2791708/4868607/2024_02_01_FCC_Iakovidis%20fullHadronic.pdf
https://pdg.lbl.gov/2023/listings/contents_listings.html

Baseline approach

- "exclusive" Durham kt algorithm for N = 4
  - determine dij between each pair of particles

    $$d_{ij} = 2min\left(E_i^2, E_j^2\right)\left(1 - cos\theta_{ij}\right)$$

  - recombine i, j pair with smallest dij, and update all distances
  - stop when you have reached a predetermined number of jets (here N = 4)
- Jet flavour tagging
  - jet tagging returns probabilities for each jet to be of given flavor: P(u), P(d), P(g), P(b), P(s),..
- Jet pairing
  - matching jets to form Higgs and Z candidates

ACKS : Michele Selvaggi, Dolores Garcia
https://arxiv.org/pdf/1111.6097 (Cacciari, Salam, Soyez)
https://arxiv.org/pdf/2202.03285 (Bedeschi, Gouskos, Selvaggi)

HZ decay mode

Off-diagonal:

e.g Z(bb)H(ss): requesting 2b-jets and two s-jets automatically “tags” the jets coming from the Higgs and the Z

vs. diagonal:

eg: Z(ss)H(ss): tagging does not help (25%)

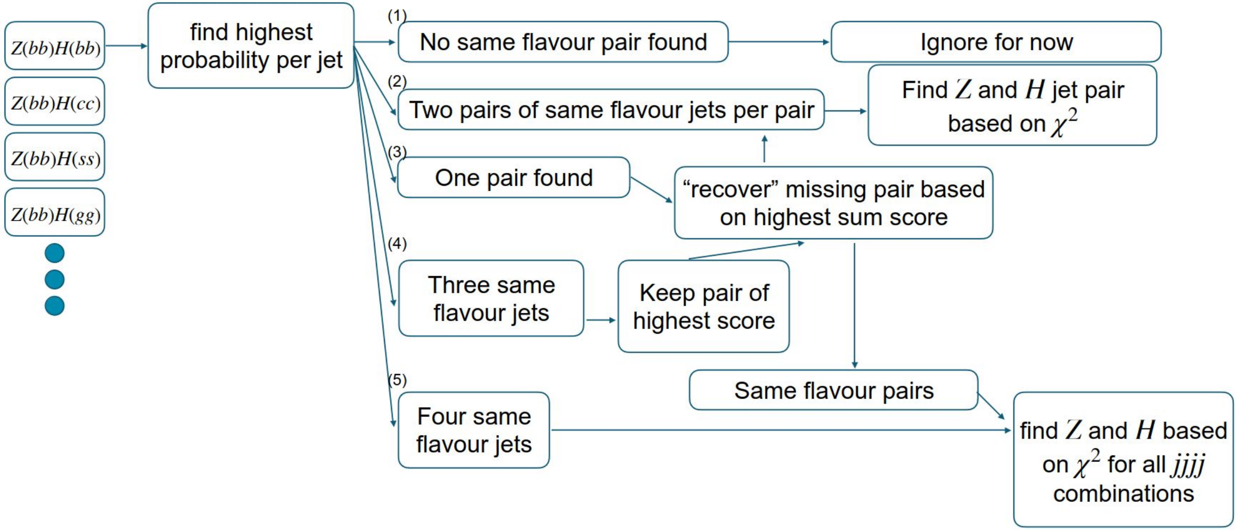
	H	bb	cc	dd	ss	uu	gg
Z	proba	58	2	0	0,02	0	8
bb	15,2	8,816	0,304	0	0,00304	0	1,216
cc	11,8	6,844	0,236	0	0,00236	0	0,944
dd	15,2	8,816	0,304	0	0,00304	0	1,216
ss	15,2	8,816	0,304	0	0,00304	0	1,216
uu	11,8	6,844	0,236	0	0,00236	0	0,944
gg	0	0	0	0	0	0	0
					1.4%		

ACKS : Michele Selvaggi, Dolores Garcia
https://indico.cern.ch/event/1307378/contributions/5724022/attachments/2791708/4868607/2024_02_01_FCC_Iakovidis%20fullHadronic.pdf
https://pdg.lbl.gov/2023/listings/contents_listings.html

Baseline approach

- "exclusive" Durham kt algorithm for N = 4
- Jet flavour tagging (P(u, d, g, b, s,..))
- Jet pairing using Z and H mass constraints

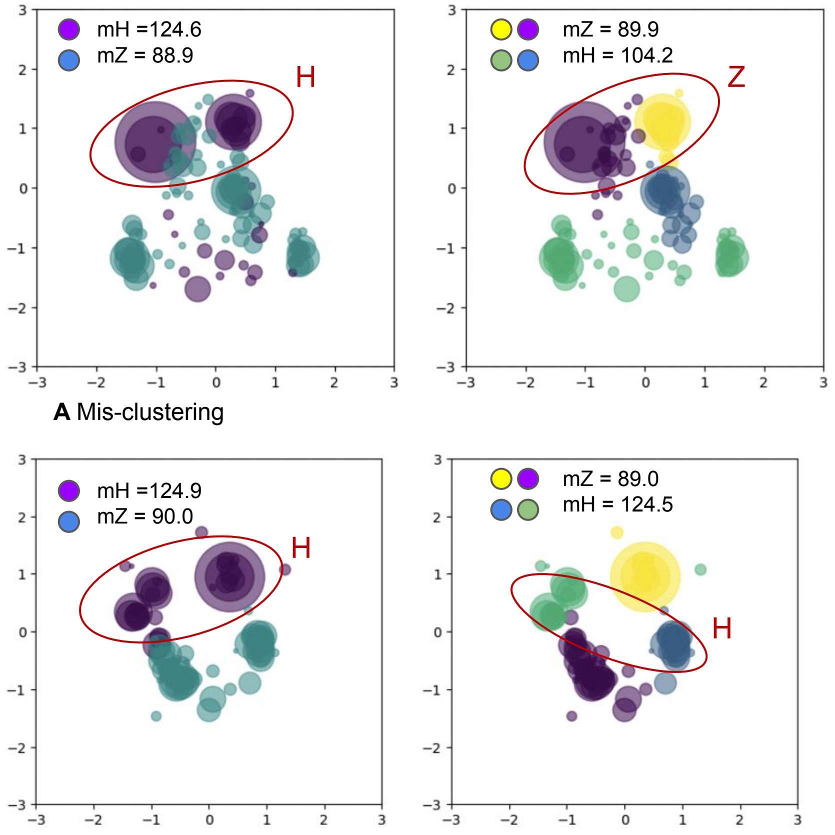
$$\chi^2 = \frac{\left(M_{\frac{1}{2}} - M_H\right)^2}{\sigma_H} + \frac{\left(M_{\frac{1}{2}} - M_Z\right)^2}{\sigma_Z}$$



Limitations

Loss in performance can be due to:

- Mis-clustering of soft particles leading to degraded resolution
- Miss matching of jets pairs



A Mis-clustering

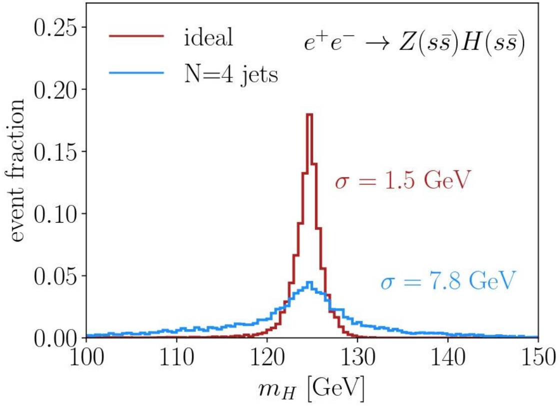
B Mis-matching of jets pairs

ACKS : Michele Selvaggi, Dolores Garcia

[1]Fujii, K., Grojean, C., Peskin, M. E., Barklow, T., Gao, Y., Kanemura, S., ... & Murayama, H. (2020). ILC study questions for snowmass 2021. *arXiv preprint arXiv:2007.03650*..
[2] Gallicchio, J., & Schwartz, M. D. (2010). Seeing in color: jet superstructure. *Physical review letters*, *105*(2), 022001.

What can we gain?

factor ~4 of improvement available



Possible solutions:

- Parameter tuning (generalise distance metric ?)
  - trial and error
- ML Learning distance metrics? piecewise continuous function, hard optimization problem
- **ML end-to-end approach**
  - **classify each final state particle as originating from Z or H**
  - **possible because $\Gamma_H \ll \Gamma_Z$**

→ Color Singlet Clustering

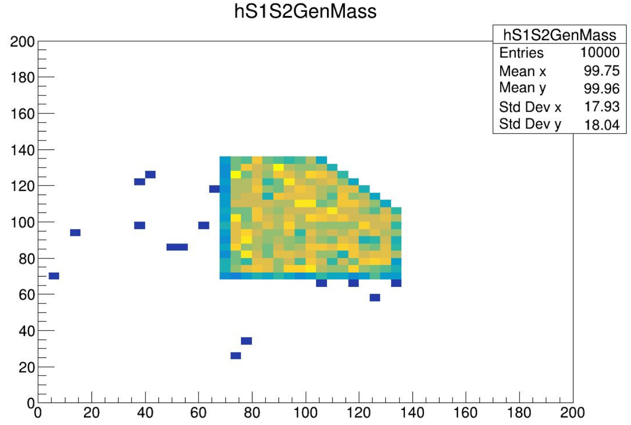
ACKS : Michele Selvaggi, Dolores Garcia

[1]Fujii, K., Grojean, C., Peskin, M. E., Barklow, T., Gao, Y., Kanemura, S., ... & Murayama, H. (2020). ILC study questions for snowmass 2021. *arXiv preprint arXiv:2007.03650*..
[2] Gallicchio, J., & Schwartz, M. D. (2010). Seeing in color: jet superstructure. *Physical review letters*, *105*(2), 022001.

Clustering Color Singlets : dataset

Our training dataset:

- Simple case of two colors singlets (S1, S2) decaying to ssbar
- Uniform mass distribution of S1 and S2 (treat more general decay than HZ)
  - avoid the network ”learning” the Higgs and Z mass
  - e.g. apply to ZZ or WW decays
- Stable final state particles from hadronisation (truth label: S1 or S2)
- No resolutions detectors effects



(px, py, pz)

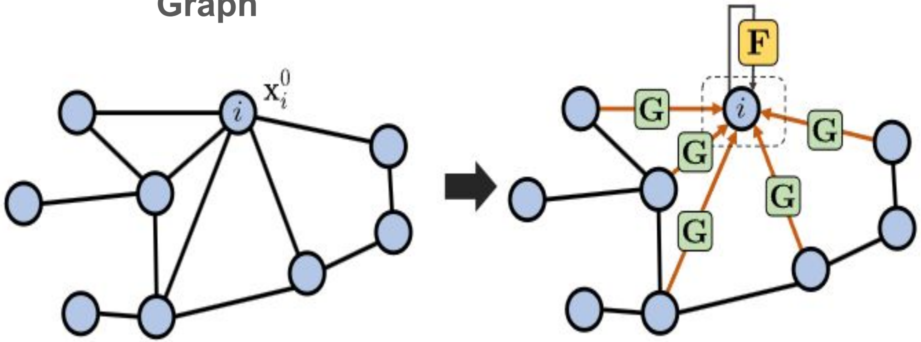
(p, theta, phi)

only use final state particle **kinematic properties** for an apple-to-apple comparison to jet clustering approach

ACKS : Michele Selvaggi, Dolores Garcia

Clustering Color Singlets : dataset

Graph



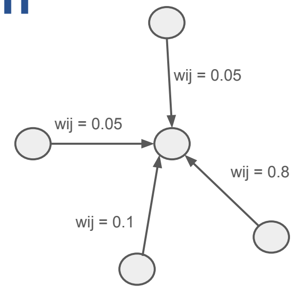
Characteristics of the graph :

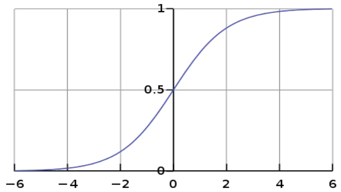
Nodes == final state particles :

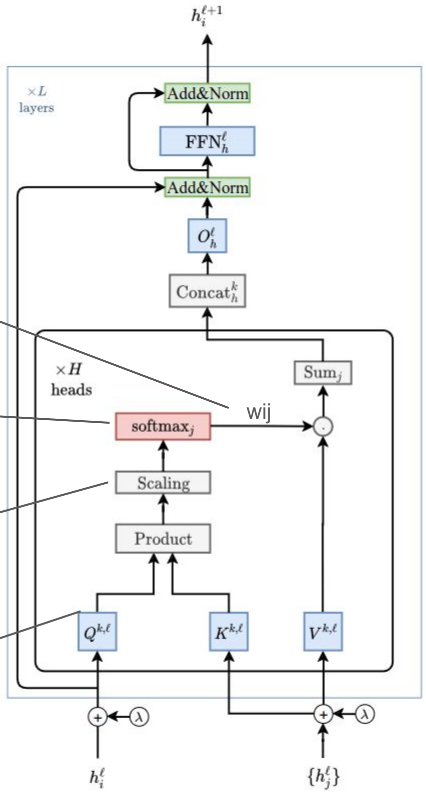
- px
- py
- pz

- Nodes == final state particles with 3 features
- 3 features embedded in higher dimensional space R3 -> RM (M>3)
- Each node does message passing: it receives information from its neighbours

Clustering Color Singlets : Graph transformer model







Graph Transformer Layer

Case with edges :

- dij (from "exclusive" Durham kt algorithm)
- e_min
- e_diff
- …

for a, b adjustable coefficients: $y = a * x + b$

Multiply the entry by a matrix with adjustable coefficients

ACKS : Michele Selvaggi, Dolores Garcia

[1]Dwivedi, V. P., & Bresson, X. (2020). A generalization of transformer networks to graphs. *arXiv preprint arXiv:2012.09699*..

Clustering Color Singlets : Graph transformer model

Figure of merit : reconstructed mH and MZ

- Cluster : baseline approach with

$$\chi^2 = \frac{\left(M_{\frac{1}{2}} - M_H\right)^2}{\sigma_H} + \frac{\left(M_{\frac{1}{2}} - M_Z\right)^2}{\sigma_Z}$$

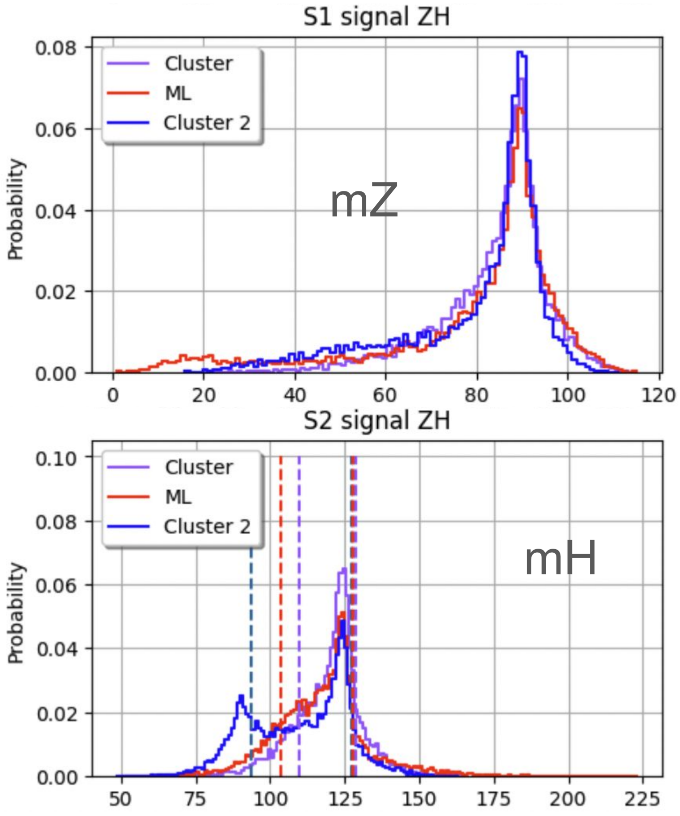
- Cluster 2 : baseline approach with

$$\chi^2 = \left(M_{\frac{1}{2}} - M_Z\right)^2$$

- ML-GNN approach

Results :

- ML is able to predict mH and mZ peaks
- However worse resolution (compared to cluster with kT exclusive clustering + jet pairing) on Higgs peak

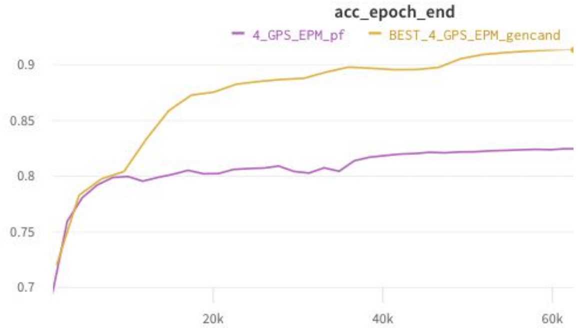


ACKS : Michele Selvaggi, Dolores Garcia

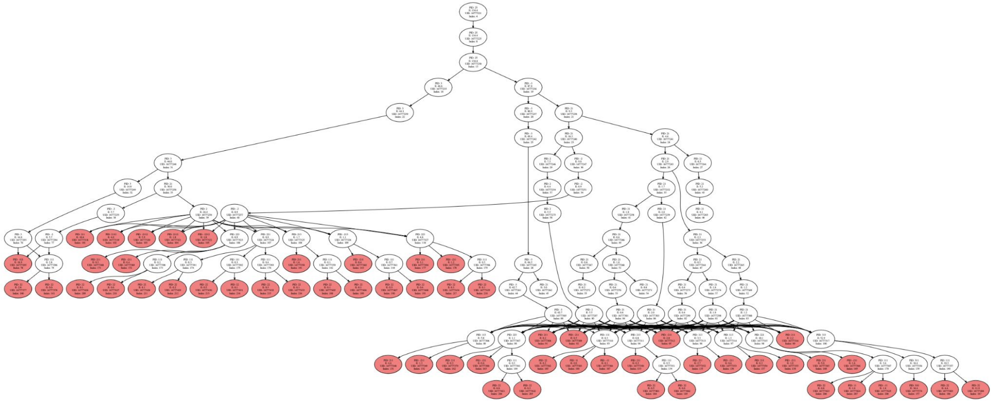
CSC- Possible improvements (I)

Could include extra information:

- **The generation tree**
  - **Only labeling final state particles to S1 and S2**
  - **Graph wiring is important**
  - Using information about **the ordering** (<tree structure) performance can be improved
  - Efforts to obtain MLE (A*, beam search…) all for small number of leaves [1,2]
- **Additional features (PID, displacement …)**



A. Accuracy increase with new wiring, ordering by tree structure



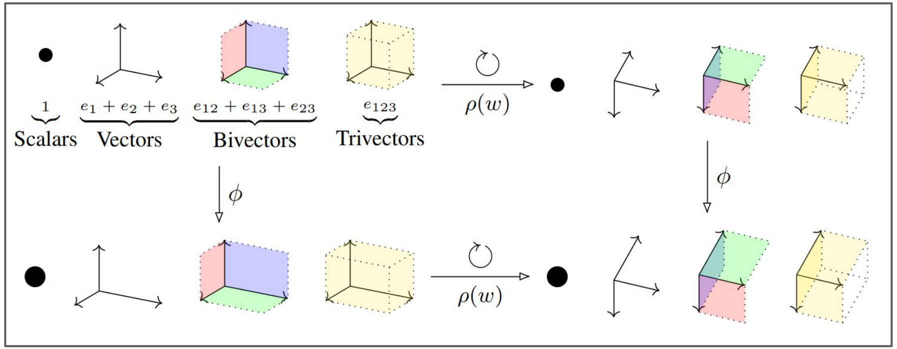
B. Example tree

ACKS : Michele Selvaggi, Dolores Garcia
[1]Brehmer, J., Macaluso, S., Pappadopulo, D., & Cranmer, K. (2020). Hierarchical clustering in particle physics through reinforcement learning. *arXiv preprint arXiv:2011.08191*.
[2] Greenberg, C. S., Macaluso, S., Monath, N., Dubey, A., Flaherty, P., Zaheer, M., ... & McCallum, A. (2021, December). Exact and approximate hierarchical clustering using A. In *Uncertainty in Artificial Intelligence* (pp. 2061-2071). PMLR.

Clustering Color Singlets - Possible improvements (I)

Advantages :

- Physical adapted
- Lorentz equivariant
- Respect Minkowski metrics



ACKS : Michele Selvaggi, Dolores Garcia
Clifford Group Equivariant Neural Networks

Conclusion

- The fully hadronic of HZ decay at FCC-ee is a challenge for the measurement of mH and mZ
- The baseline approach (“exclusive” Durham kt algorithm, jet flavour tagging, jet pairing) has limitations (mis-clustering, mis-pairing)
- Overcome them with end-to-end GNN approach for color singlet clustering
- Promising results for the Graph transformer model
- Working on new ideas

ACKS : Michele Selvaggi, Dolores Garcia
Clifford Group Equivariant Neural Networks

Thank you!