# Self-supervised learning of Lorentz invariant models

## ML4Jets 2026, Vienna

Andreas Hermansen[1] Tobias Golling[1] Slava Voloshynovskiy[2]

[1] Université de Genève, Département de physique nucléaire et corpusculaire
[2] Université de Genève, Centre Universitaire d'Informatique

andreas.hermansen@unige.ch

2026-09-16

# Advances in Jet Tagging

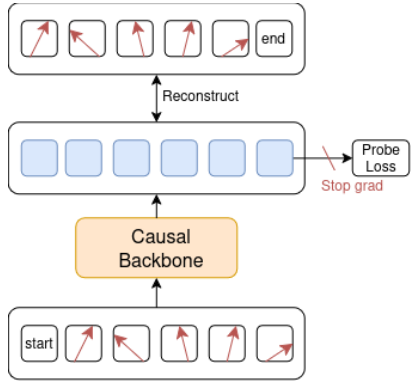


- Success of Lorentz Invariant (LI) models
  - PELICAN, ParT, L-GATr(-slim), etc.
- Success of scaling / self-supervised learning (SSL)
  - Masked Particle Modeling (MPM), Omnijet-$\alpha$, Omnilearned, etc.
- Question: How to combine the two?

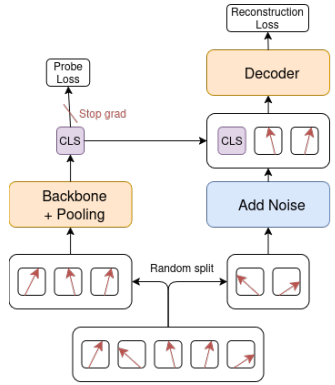
# Challenges of SSL for LI models

## Masked-generative pretraining

- Auto-Regressive modelling / Masked token prediction
- Converting masked prediction to LI
  - Tokenization breaks LI
  - Large numerical range of $E, p_T$
    - Log transforms break LI




NTP Transformer

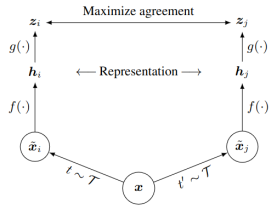



Set-to-Set Flow Matching

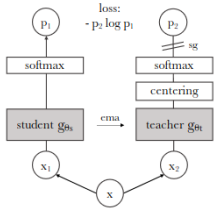
# Joint embedding pretraining



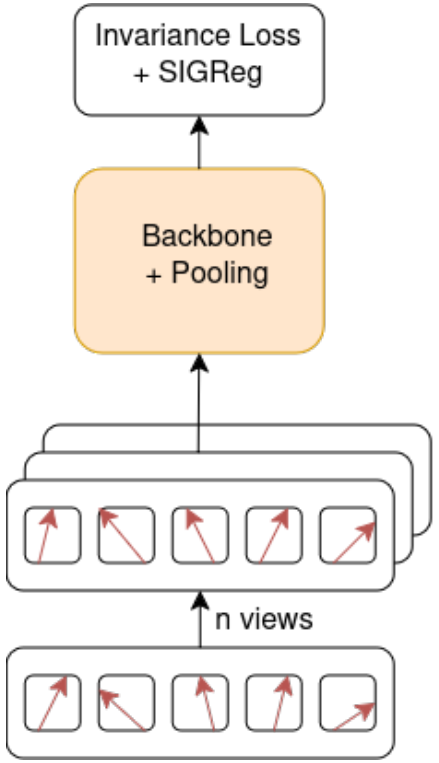
- Generative Pretraining - challenges
  - Mostly successfull in NLP
  - Initial attempts in CV (MAE / Diffusion MAE)
- Recent SSL in CV purely Joint Embedding
  - Contrastive, DINO, JEPA, etc.
  - Pixel-Level too finegrained




SimCLR[1]




DINO[2]



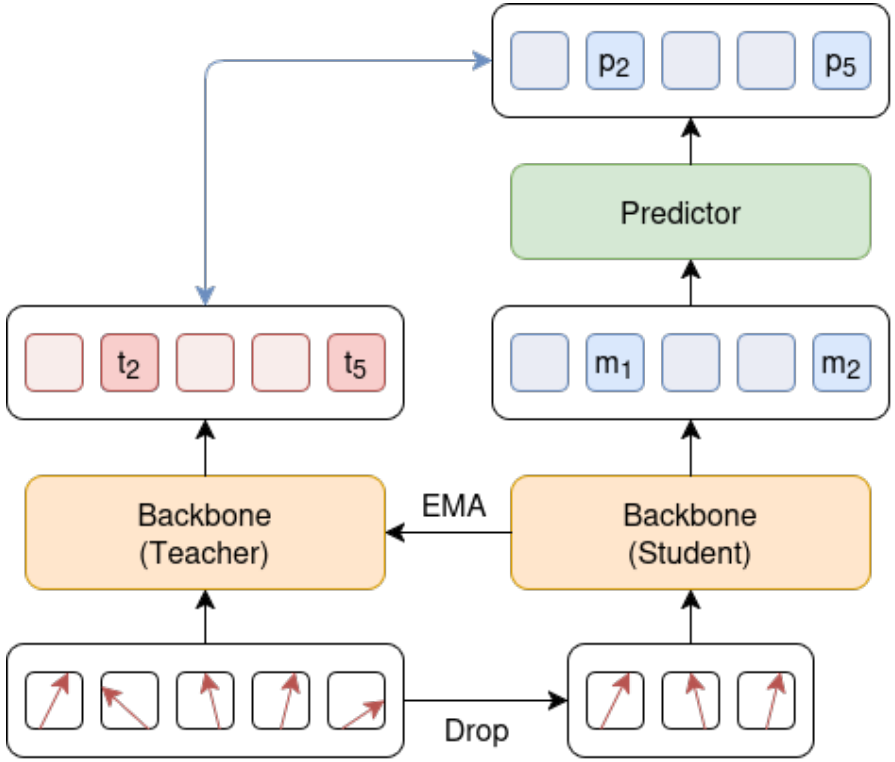

LeJEPA

[1]Figure 2 from Chen et al. [arXiv:2002.05709]

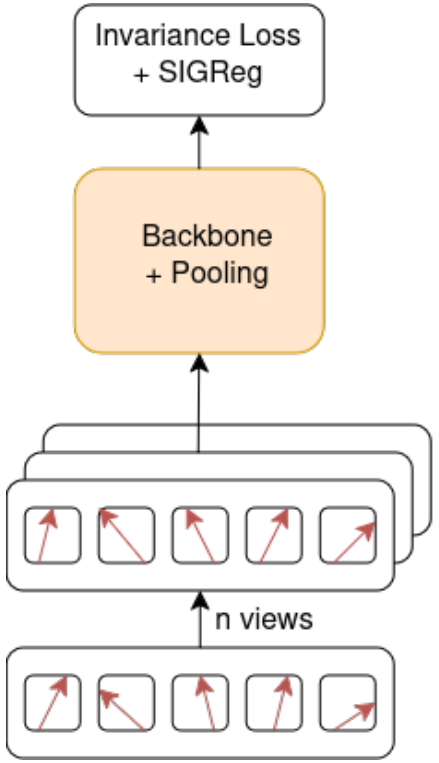
[2]Figure 2 from Caron et al. [arxiv.org/abs/2104.14294]

# Evolution of JEPA's





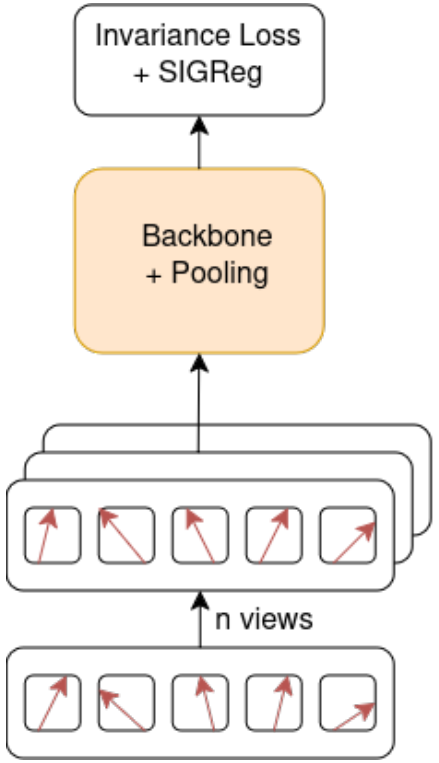

Previous iterations of JEPA




LeJEPA

# LeJEPA

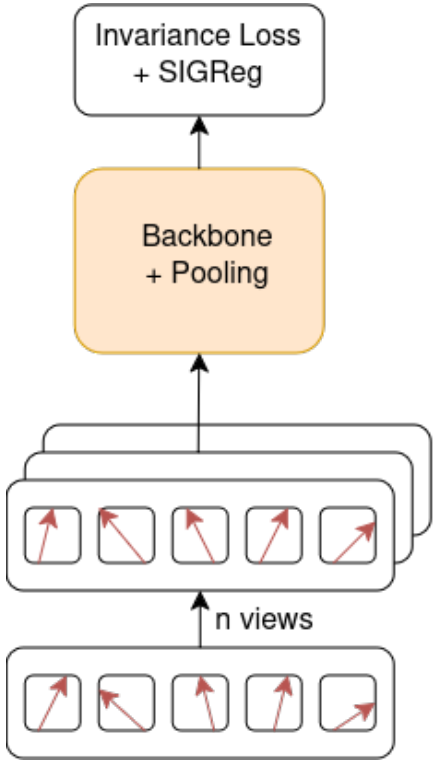
- Greatly simplifies design
  - No Student-Teacher
  - No EMA
  - No Stop-grad
- Invariance loss
  - Views = same embedding
    - $f_\theta(A(x)) \approx f_\theta(x)$
  - SIGReg
    - Make latent space Gaussian




LeJEPA

# LeJEPA: How to generate views?

- Views = Augmentations
- Based on JetCLR
  - arxiv.org/abs/2108.04253
  - Colinear splitting
  - Smearing $\eta$ / $\varphi$
  - Subsampling (local view)
  - SO(1, 3)
- Free to include more




LeJEPA

# Why Lorentz invariance



- Physics invariant under $\mathrm{SO}^+(1,3)$
  - Beam axis reduces to $\mathrm{SO}^+(1,1) \times \mathrm{O}(2)$
- Lorentz invariant networks
  - $f_\theta(\rho(\Lambda)x) = f_\theta(x)$ exactly
  - restrict functionclass - regularization
- LeJEPA: the same invariance, learned
  - $f_\theta(A(x)) \approx f_\theta(x)$ over views $A$
    - $A$ can capture more than $\mathrm{SO}^+(1,3)$
      - Detector smearing
      - Particle Physics knowledge

# Transformer to L-GATr-slim

| | Transformer | L-GATr-slim[1] |
|---|---|---|
| Token | $x \in \mathbb{R}^d$ | $(s, v) \in \left(\mathbb{R}^{c_s} \times \mathbb{R}^{c_v \times 4}\right)$ |
| Linear | $Wx + b$ | $(W_s s + b_s, W_v v)$ |
| Norm | $\left(\sum_i x_i^2\right)^{-\frac{1}{2}}$ | $\left(\sum_c \lvert\langle v_c, v_c\rangle\rvert^2 + \sum_c s_c^2\right)^{-\frac{1}{2}}$ |
| attention logit | $q \cdot k$ | $q_s \cdot k_s + \sum_c \langle q_{v,c}, k_{v,c}\rangle$ |
| FFN (GEGLU) | $\mathrm{GELU}(W_1 x) \odot W_2 x$ | scalars: as left;<br>vectors: $\mathrm{GELU}(\langle W_1 v, W_2 v\rangle) \odot W_3 v$ |

Same residual blocks; per-channel factors and $\varepsilon$ omitted. Attention logit normalisation not included.

[1]See also Antoine Petitjean's & Jonas Spinner's paper arxiv.org/abs/2512.17011 and talks on: Economical Jet Taggers -- Equivariant, Slim and Quantized & Virtues and Vices of Equivariant Transformers

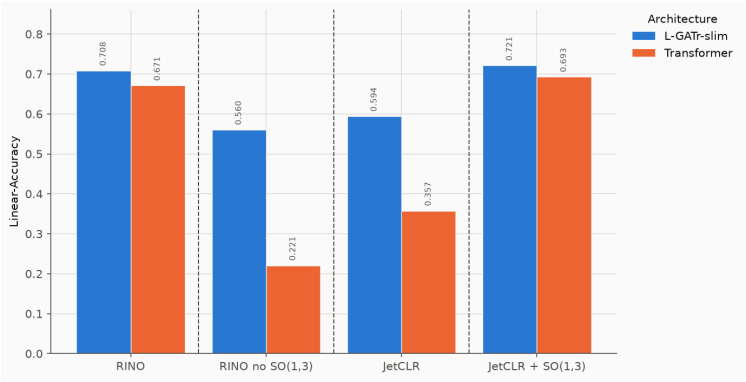
# Broken symmetries

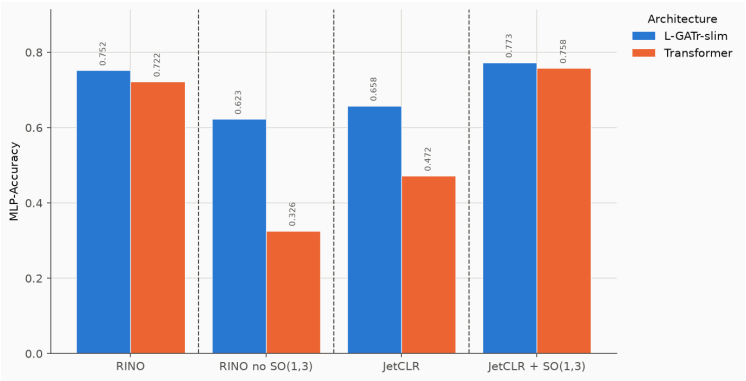
- L-GATr-slim: $\mathrm{O}(1,3)$[1] invariant.
  - Introduce spurions:
    - Beam: $\mathrm{O}(1,3) \to \mathrm{O}(1,1) \times \mathrm{O}(2)$
      - Boost along axis, rotations around axis
    - Time: $\mathrm{O}(1,3) \to \mathrm{O}(3)$
      - Remove boosts
    - Both: $\mathrm{O}(1,3) \to \mathrm{O}(2)$
      - Rotations around beam

[1] Similar reductions follow if initial symmetry is chosen as $\mathrm{SO}(1,3)$ / $\mathrm{SO}^+(1,3)$

# Probes: Augmentations

- Why probes?
  - Measures information extraction of pretraining
  - Latent calibration
- L-GATr-slim outperforms Transformer across augmentations
- JetCLR[1] vs. RINO[2]
  - Introducing full $SO(1,3)$ improves probes.



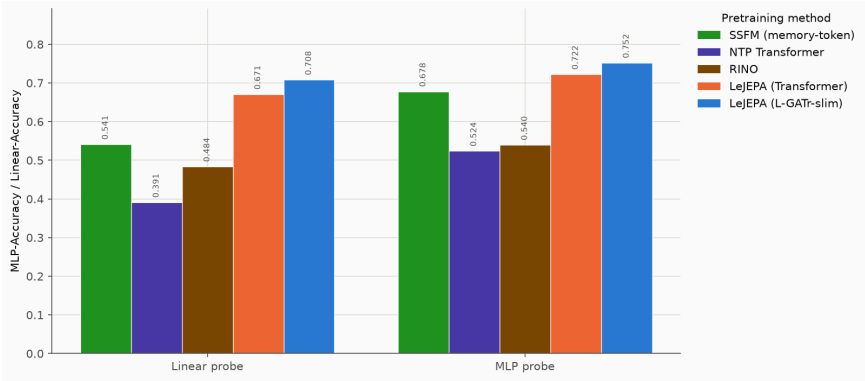




[1] arxiv.org/abs/2108.04253

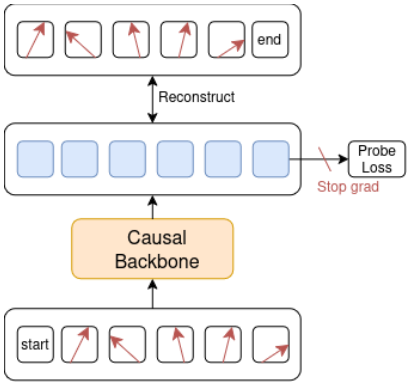
[2] arxiv.org/abs/2509.07486

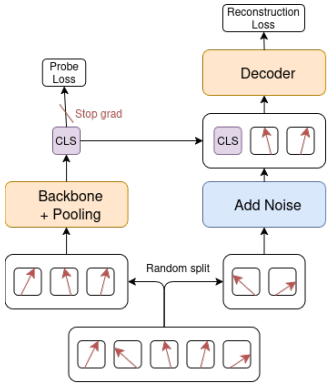
# Probes: SSL methods

- LeJEPA achieves best separation
- Compared against
  - NTP-Transformer (AR)
  - Set-to-Set Flow Matching (Diffusion MAE)
  - RINO (DINO inspired model)



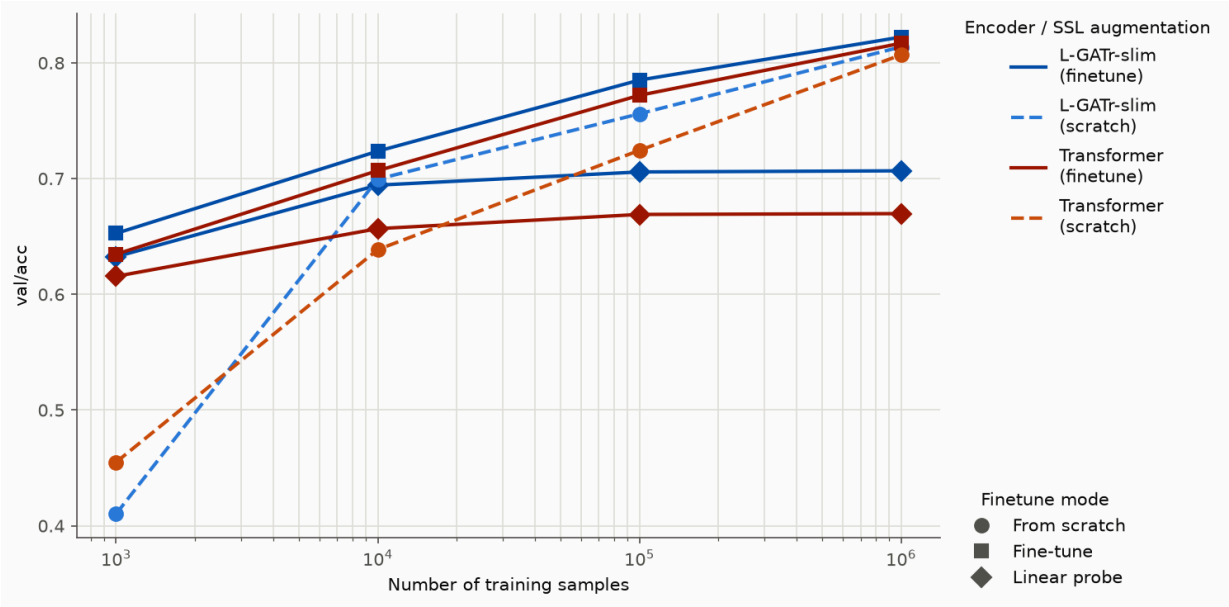

Probe accuracy during pretraining.



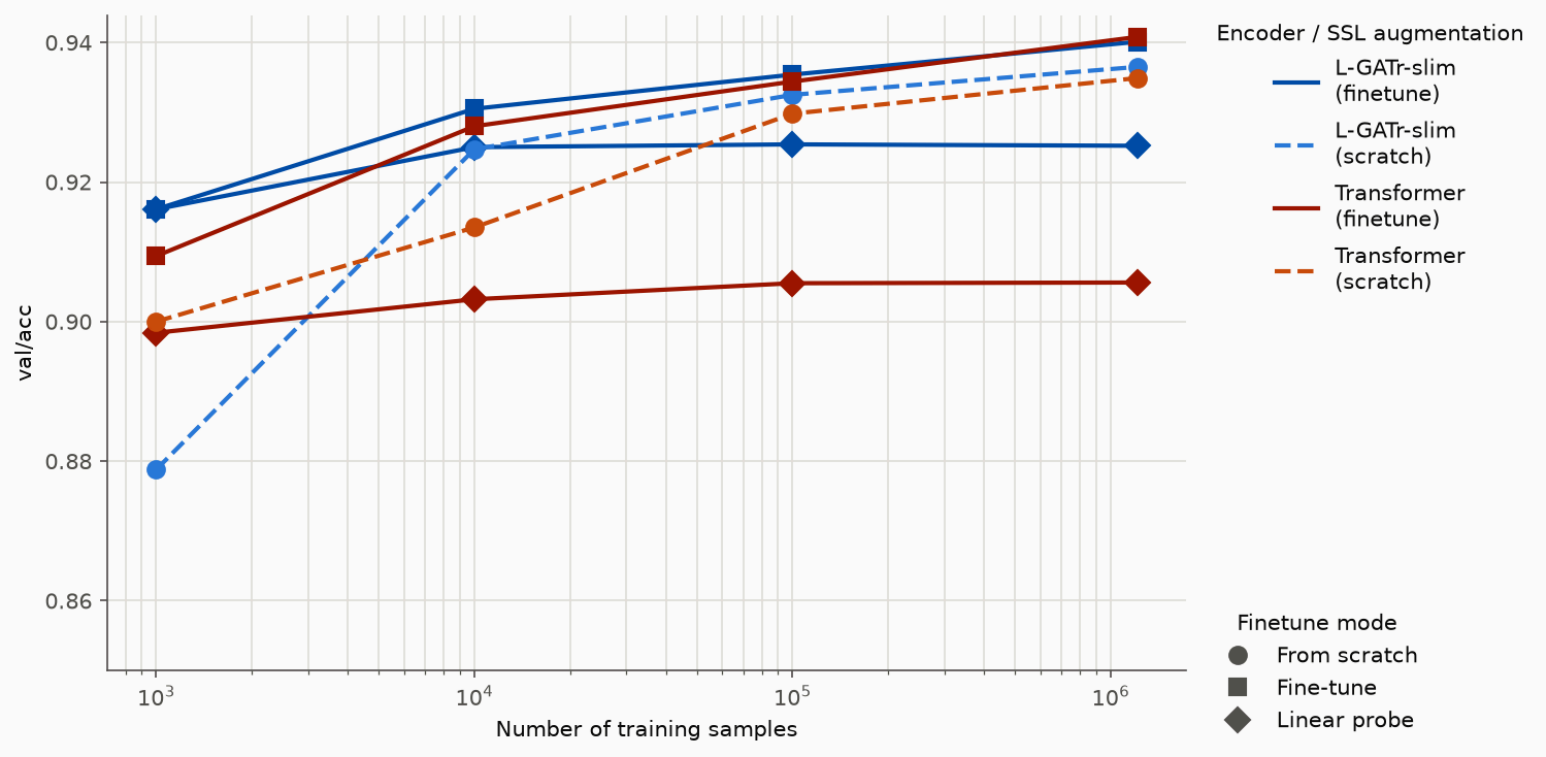
# Data size: classification (In Domain)

- Pretraining helps
  - Few samples
- Symmetry helps
- Can be combined



# Data size: cross-detector transfer

- Similar story to in-domain
- CMS to ATLAS



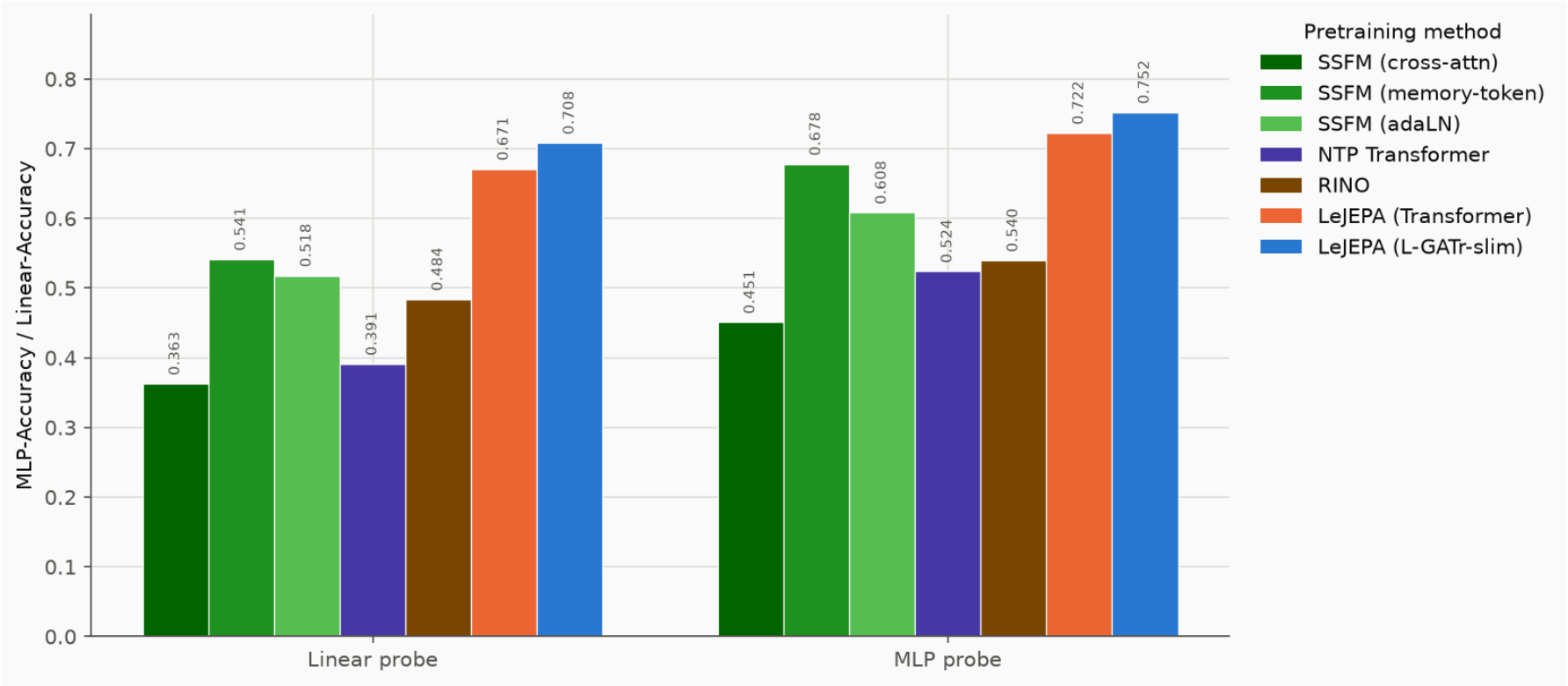
# Summary and outlook

- Joint embedding pretraining strong alternative to generative pretraining
- LeJEPA achieves remarkable Linear/MLP probe performance
  - Mitigate domain shift
  - Calibrate latent representation
- Self-supervised pretraining of LI models possible
  - Improvements from LI architectures and SSL not mutually exclusive
- How does this behave at scale?

# Backup Slides

# Backup: Full probe comparison






All pretraining methods.

# Backup: SIGReg

- (Hyper-spherical) Cramer-Wold theorem
  - $X, Y$ Random variables in $\mathbb{R}^n$

$X \overset{d}{=} Y$ iff. $\langle a, X \rangle \overset{d}{=} \langle a, Y \rangle$ for all $a \in \mathbb{S}^{n-1}$

## Test-statistic as loss

$A$: Collection of unitvectors: $A = \left\{ a_i \in \mathbb{S}^{n-1} \right\}_{i=1}^{K}$

$$\text{SIGReg}_T\left(A, \left\{ f_\theta(x_n) \right\}_{n=1}^{N}\right) = \frac{1}{|A|} \sum_{a \in A} T\left(\left\{ \langle a, f_\theta(x_n) \rangle \right\}_{n=1}^{N}\right)$$

- $T$ A Test statistic (Epps-Pulley)
  - $T = 0$ high p-value samples follow target distribution
  - $T \gg 0$ low p-value samples don't follow target distribution

# Backup: Epps-Pulley

$$\mathrm{EP}\left(\{x_n\}_{n=1}^{N}\right) = N \int_{-\infty}^{\infty} \|\varphi(t) - \hat{\varphi}(t)\|_2 w(t)\,\mathrm{d}t$$

- $\varphi(t)$: Target Characteristic function
- $\hat{\varphi}(t)$: Empirical Characteristic function: $\hat{\varphi}(t) = \frac{1}{N}\sum_{j=1}^{N} \exp\left(itx_j\right)$

Properties of Epps-Pulley:

- Well behaved gradient and second derivative