ML4Jets 2026

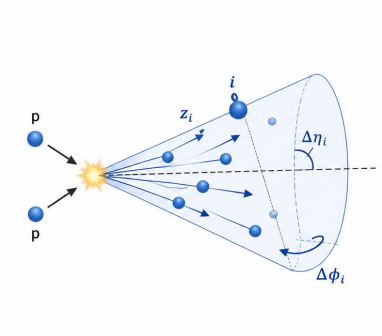
Towards Foundation Models for Heavy- and Light-Ion Collisions

João A. Gonçalves

Lucie Flek Matthias Schott

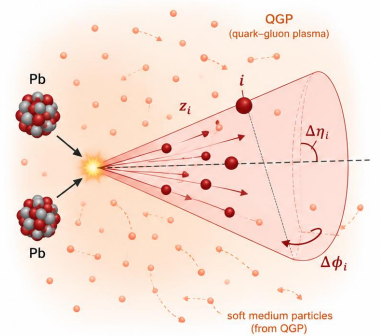
Vienna 14 September 2026

Introduction



pp: clean shower

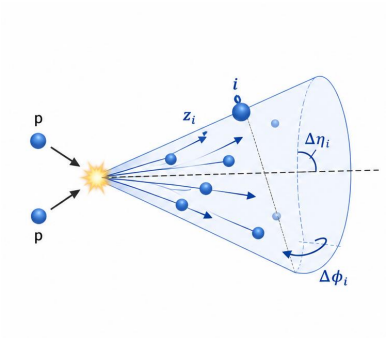
QCD
across scales



PbPb: modified shower and bulk information

The physics spans the whole collision, not only the jet.

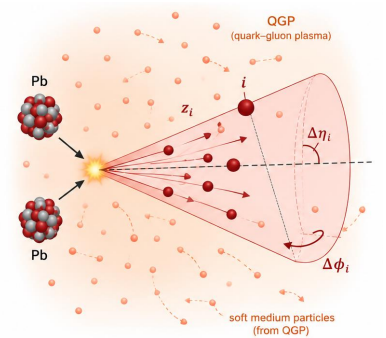
Motivation



pp: clean shower

$\neq$

not pp with
more particles

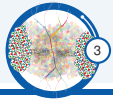


PbPb: modified shower and bulk information

Heavy- and light-ion foundation models warrant study and benchmarking independently of pp foundation models.

Of course, we would like an LHC-wide collider foundation model.

Dataset



CMS 2010 PbPb Open Data, $\sqrt{s_{\mathrm{NN}}} = 2.76$ TeV

CMS record 14011; CMS particle flow.

Recorded collision data
151.0 TiB

→

PF Reconstruction
≈1.52 TB

→

Particle selection
Finite features
$p_{\mathrm{T}} > 1$ GeV; $|\eta| < 1$

→

Processed particles
46.11M nonempty records
9.30B particles; 181.2 GB

1M benchmark

1,000,000 / 250,000 / 250,000

Train / validation / test; fixed record indices.

Full-corpus campaign

44,107,877 / 1,000,000 / 1,000,000

Train / validation / test.

Stratified by multiplicity, event kinematics and PF composition.

Store all selected particles; training uses at most the 1,000 highest-p_{T} per event.

Data Representation

Particle representation

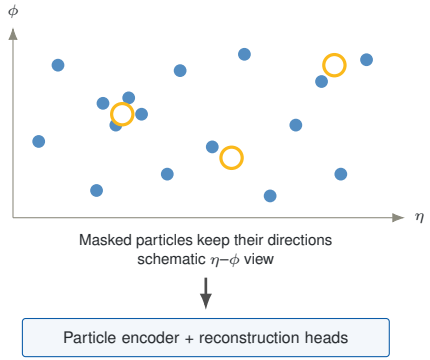
$$x_i = \left(\log(1 + p_{\mathrm{T}}), \eta, \sin\phi, \cos\phi, \log(1 + m), c_i\right)$$

c_i: categorical particle-flow (PF) identity.

Particle masking

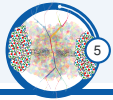
Mask 15% of particles; recover p_{T}, mass and PF identity from event context.

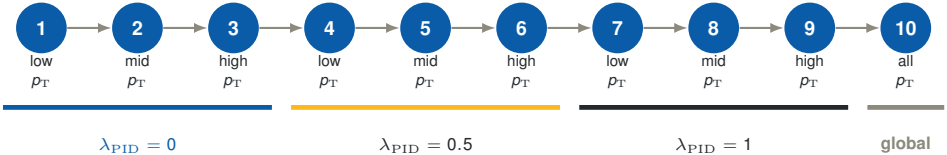
Directions (η, ϕ) remain visible.



$$\mathcal{L} = \tfrac{1}{2}\left(\mathrm{MSE}_{\log(1+p_{\mathrm{T}})} + \mathrm{MSE}_{\log(1+m)}\right) + \lambda_{\mathrm{PID}}\,\mathrm{CE}_{\mathrm{PF\ PID}}$$

Training Strategy





Kinematics, then identity

Mask 15% of the whole event, sampling within the active p_T third.

Final stage: mask across all p_T and lower the learning rate.

Batching and early stopping

Order batches by particle multiplicity to manage complexity increase in training stages.

Validation-based early stopping controls stage length.

Very initial training curriculum proposal, bounded by computational resources.

Models Considered

Collider architectures

ParT / MIParT
Particle-pair interactions

OmniLearned
Local + global attention

HI: random → PbPb.
pp-HI: public pp → PbPb.

Spatial representations

Concerto
3D point-transformer

Adapted to particle coordinates
Geometric structure beyond a sequence

Not collider-specific.

Sequence architectures

ModernGBERT: bidirectional attention

Causal Qwen: recurrence + causal attention

Mamba-2: state-space mixing

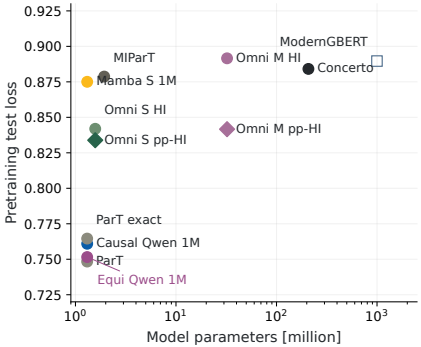
Equi Qwen: separate, position-free attention.

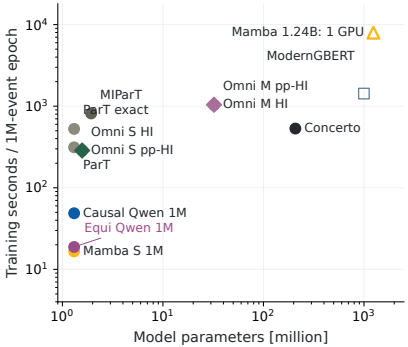
Common particles, masking and targets.

Compare architecture, training cost and initialization separately.

Models Considered

Size, reconstruction and epoch cost

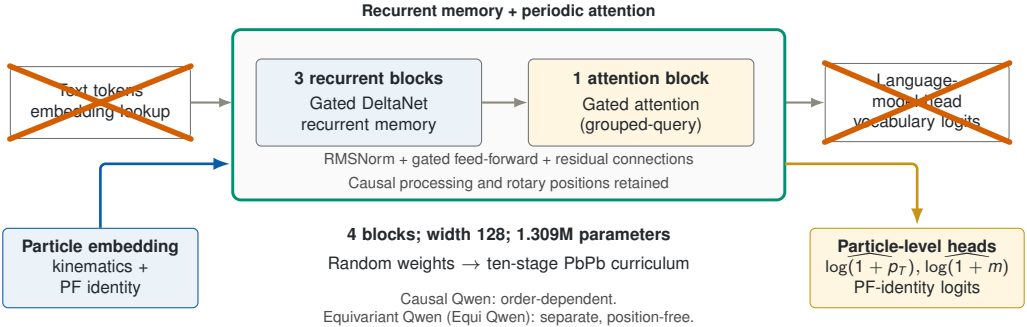




Similar size can mean very different training cost and pretraining loss.

Original Model

3. Add particle inputs and heads



MoE not used yet — possible extension
Experts per collision system, centrality bin or hadronic-calorimeter activity bin (to test).

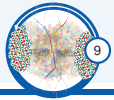
Qwen3.5 architecture; particle-adaptation schematic.

Both Qwen models here are dense: no experts or router.

Reuse the architecture; learn the weights from PbPb data.

Downstream Benchmark

Four transfer tasks; broader physics next



Current benchmark

Four simulation-based tests

Five-class particle identity

Matched versus unmatched jets

Signal-generator jet p_{T}

JEWEL: vacuum versus medium

Fine-tuning versus scratch.

Results on all four tasks

84 completed runs:
53 fine-tuned; 31 scratch.
JEWEL public sample; published comparison.

Dream benchmark

Hard and soft probes

Heavy flavour and quenching

Flow, geometry and correlations

Unfolding and medium properties

New tasks need fixed selections and validated reference measurements.

OmniFold-HI; flow correlations.

Attainable benchmark

CMS PbPb, pPb and pp
Collision data + simulation

ATLAS minimum-bias PbPb
221M events + separate simulation

CMS releases; ATLAS PbPb + MC.

Downstream Tasks

Particle identity

Reconstructed particles $\rightarrow$ five generator-matched classes.

PF category is input; not raw-detector PID.

Hard-jet tagging

Generator-matched versus unmatched jets.

Anti-k_t, $R = 0.4$; ROC and rejection.

Signal jet momentum

Reconstructed event $\rightarrow$ clean signal-generator jet p_{T}.

Embedded simulation; not pre-quenching momentum.

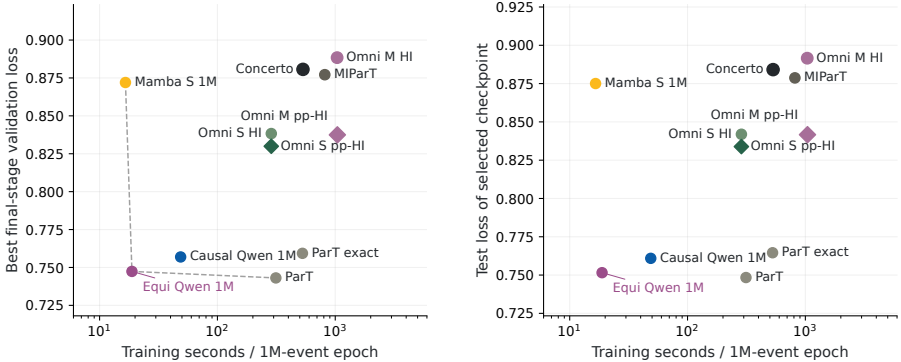
JEWEL medium tagging

Vacuum versus medium jets in JEWEL.

Medium response + underlying event; native weights.

One task per run; four initialization routes

Scratch FT from PbPb PT FT from pp PT FT from pp $\rightarrow$ PbPb PT



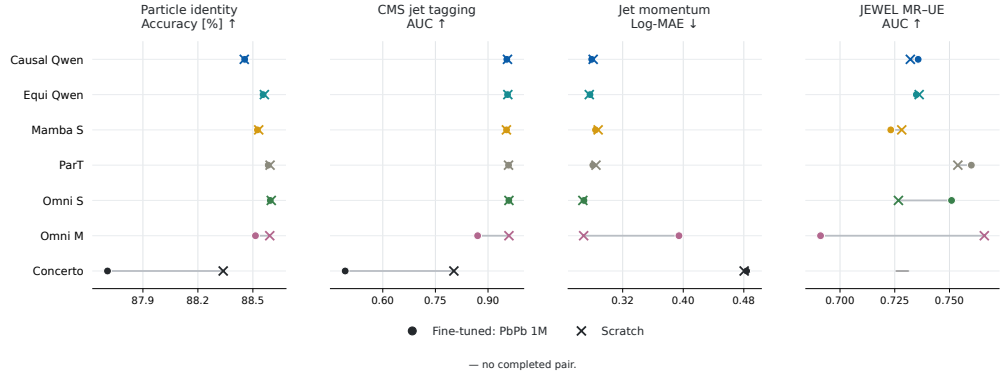
1M cohort, 4 H200s. Selection uses validation; test is held out. Dashed: observed validation trade-off.

Four $\simeq$1.31M encoders: 120–151 epochs on four H200s.
Final-stage training median: 17–313 s/epoch. Different batches/implementations; not matched compute.

Long curricula are feasible. Size alone does not determine cost.

Results

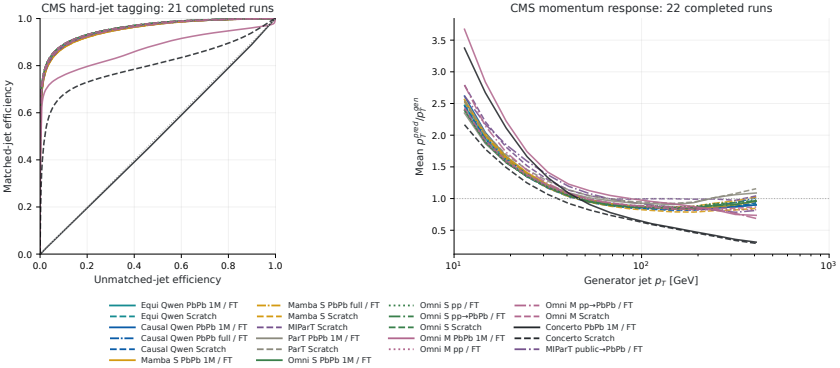
2. Does 1M PbPb pretraining beat scratch?



PbPb 1M PT → FT versus matched scratch: seven models, 27 task pairs.
Single runs. Other PT routes and full tables in backup.

Scratch is competitive. Transfer gains depend on model and task.

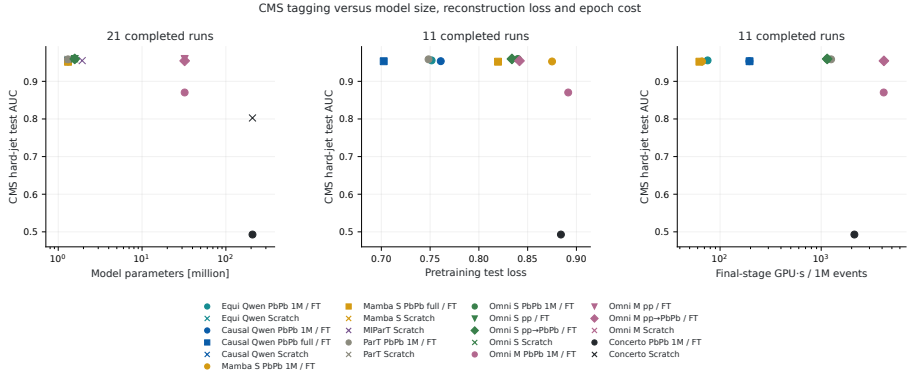
Further testing needed with larger datasets and looser kinematic cuts.



All completed independent-task runs; shared scratch once per task. Response bins require ≥100 jets.

CMS simulation test curves; FT / scratch; two reconstructed jet collections pooled.

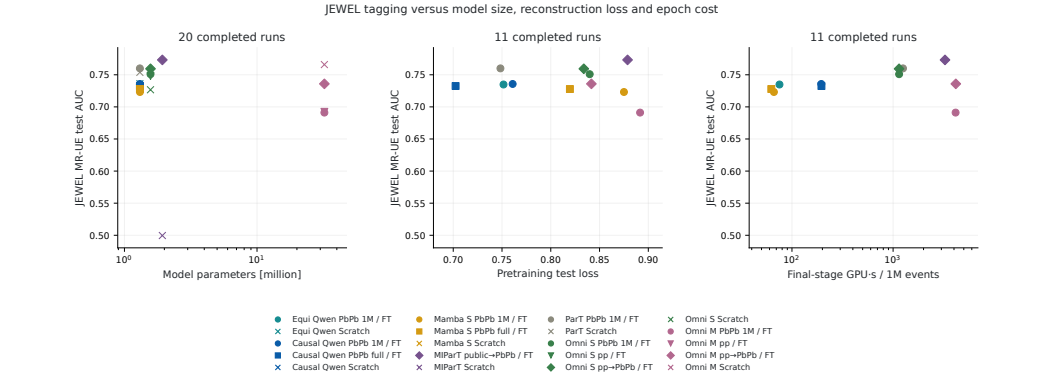
Jets can be separated; momentum response still needs calibration.



Cost: one final-stage epoch, not total pretraining. Scratch/direct pp: local PT loss and cost unavailable.

Checkpoint-matched PT loss and cost. Cost: one final-stage epoch per 1M events, not total pretraining.
PT test sets, batches and schedules differ.

Larger models and lower PT loss do not guarantee better tagging.

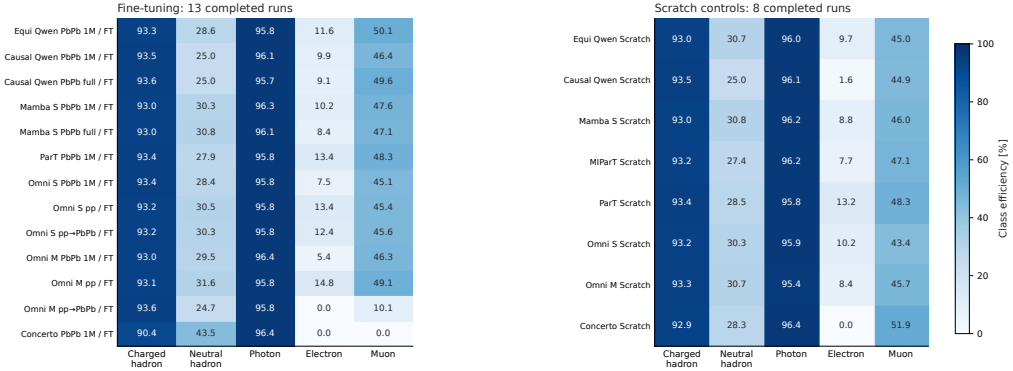


JEWEL vacuum/medium test; checkpoint-matched PT loss and cost.
Cost: one final-stage epoch per 1M events, not total pretraining. Single runs.

JEWEL transfer gains depend on the model and pretraining route.

Results

Particle identity: class recall

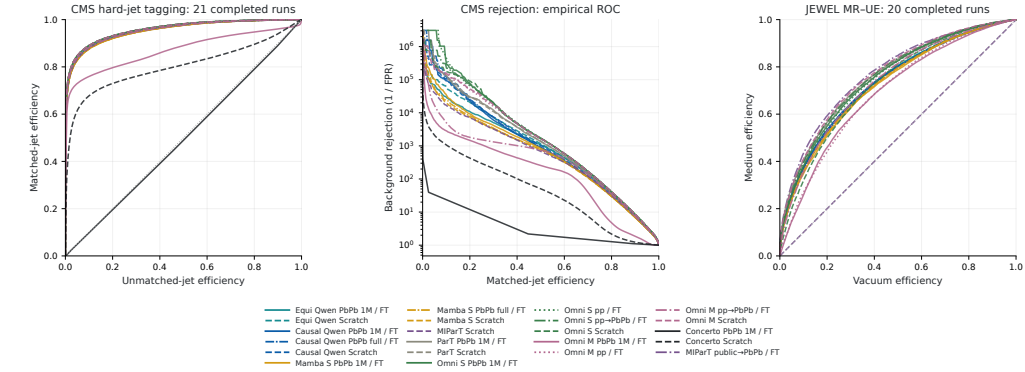


Five generator classes; reconstructed PF category remains an input. Every completed PID run is shown.

Generator-class recall [%]; PF category remains an input. All completed FT and scratch runs.

Electron and muon recall remain weak.

Possible cause: too few electrons and muons in masking. Needs verification.



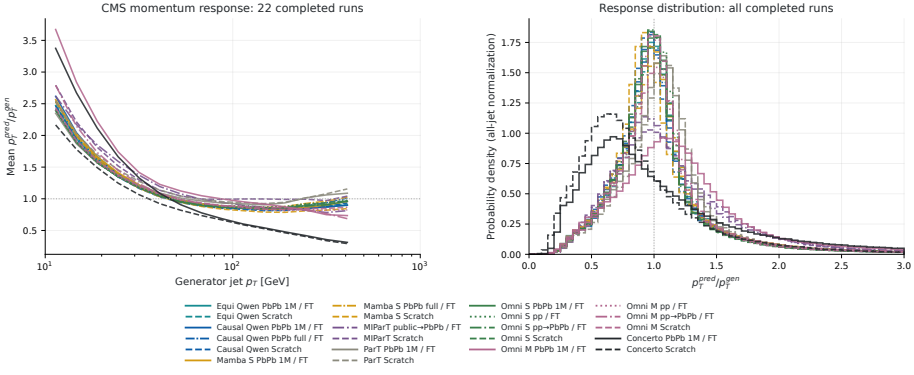
CMS rejection: finite empirical 1/FPR; scalar tables use count smoothing. Separate CMS and JEWEL tests.

CMS: matched/unmatched jets. JEWEL: a separate vacuum/medium test. Single runs.

Fine-tuning helps some models; scratch stays competitive.

Results

Signal jet momentum: fine-tuning versus scratch



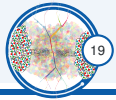
Profiles: 10–1000 GeV, ≥100 jets/bin. Density shown over 0–3; full tails retained in metrics. No uncertainty bands.

Clean signal-generator jet target; not pre-quenching momentum.
Response range 0–3 omits 1.4–5.0% of jets; metrics use full tails.

Bias: -8.4% to $+35.1\%$. Calibration remains model-dependent.

Future Work

What we have, what we would like, what we might get



Resources we have

8 H200s

BAF: up to eight concurrently.

Shared Marvin access

128 A100s + 192 A40s installed.

Shared, not dedicated; queue limits apply.

Marvin hardware; access limits.

Resources I would like to have

>100,000 GPUs

Astra: reported training scale.

A little room to explore

Models, data and physics tasks.

Aspirational, not our compute requirement.

Greg Brockman, 4 Sep 2026.

GPU type not specified in this source.

Resources we might be able to get

Tier-0 or Tier-1 cluster

A dedicated project allocation.

Better tests, then scaling

Tuned controls and repeated seeds.

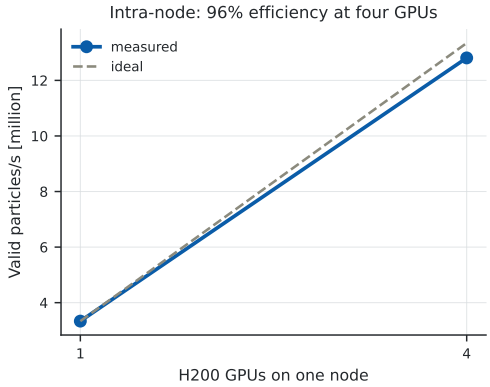
More data and larger models.

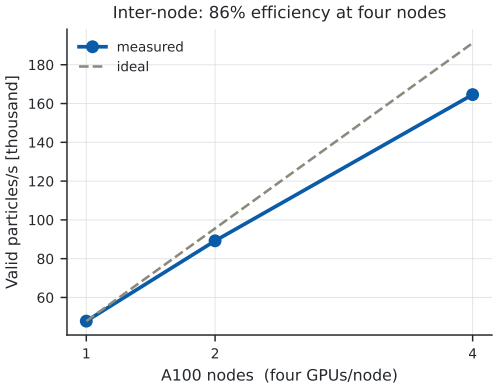
Prospective; not an awarded allocation.

More compute should buy stronger tests, not only larger models.

Future Work

Tier-0 or Tier-1 cluster: measure before scaling

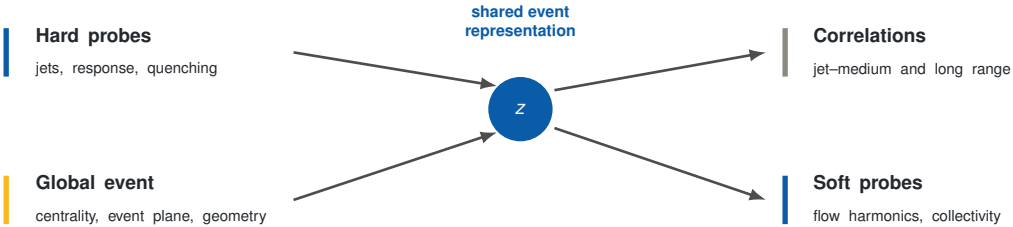




H200 / A100; fixed local batch.
Preliminary timings, especially between nodes.

Measure throughput, memory and communication before scaling.

Future Work



A common benchmark across LHC environments

pp, heavy-ion, light-ion and other collision environments.

Shared tasks, splits, reference methods and uncertainty estimates.

Cross-environment transfer remains to be tested.

Take-away

Three takeaways from the 1M benchmark

Training works

Long small-model curricula complete on four H200s.

Similar size, very different cost.

Scratch stays competitive

At 1M, fine-tuning helps only some model–task pairs.

No single model wins

Lower pretraining loss does not guarantee better downstream results.

Still to resolve: lepton recall and momentum calibration.

Limited compute: no systematic downstream HPO or robust downstream uncertainty estimates.

Does transfer improve with more data and larger models?

Test with tuned controls, repeated seeds and matched compute.

Thank you

João A. Gonçalves

jagoncalves@uni-bonn.de

Questions?

João A. Gonçalves

jagoncalves@uni-bonn.de

Open to new projects and collaborations

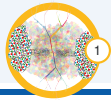
I am currently looking for my next **postdoctoral position**, working across **QCD, heavy-ion physics, and astrophysics** and **machine learning**.

Interested in collaborating or working together?

Let us talk.

João A. Gonçalves · jagoncalves@uni-bonn.de

Backup index

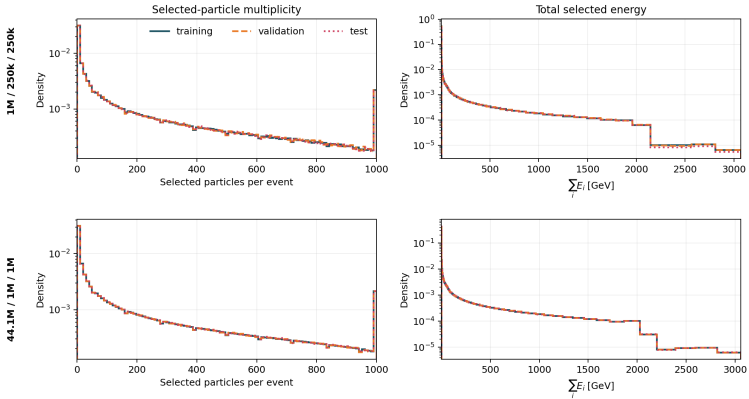


Data and objective

← Backup index

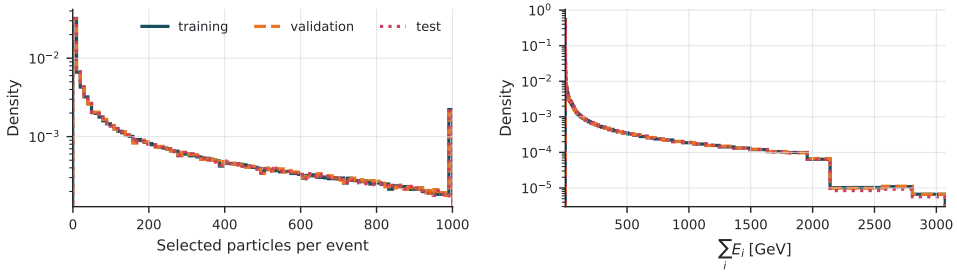
Split checks: particle multiplicity and selected energy

Split agreement for $p_T > 1\,\text{GeV}$, $|\eta| < 1$, top 1,000 particles



Dataset: Benchmark Split

Benchmark split: 1M training / 250k validation / 250k test



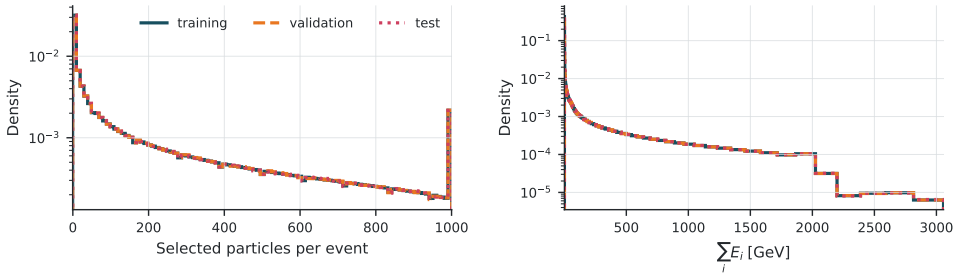
$p_T > 1$ GeV, $|\eta| < 1$, top 1,000 particles. Deterministic event split, seed 12345.

1M / 250k / 250k training, validation and test; $p_T > 1$ GeV, $|\eta| < 1$, top 1,000 particles.

Energy is in physical units, with a linear horizontal axis and logarithmic density axis.

Dataset: Full-Corpus Split

Full campaign: 44.1M training / 1M validation / 1M test

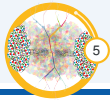


$p_T > 1\,\text{GeV}$, $|\eta| < 1$, top 1,000 particles. Independent validation and locked test splits.

44.1M / 1M / 1M training, validation and test; same particle selection.

Descriptor agreement is not proof of unique event identities or jet-level balance. Those controls remain distinct.

Exact masked-reconstruction objective



$$\mathcal{L}_{\text{cont}} = \frac{1}{2|M|} \sum_{i \in M} \left[\left(\log(\widehat{1 + p_{\text{T},i}}) - \log(1 + p_{\text{T},i}) \right)^2 + \left(\log(\widehat{1 + m_i}) - \log(1 + m_i) \right)^2 \right]$$

$$\mathcal{L} = \mathcal{L}_{\text{cont}} + \lambda_{\text{PID}} \frac{1}{|M|} \sum_{i \in M} \text{CE}(\hat{c}_i, c_i)$$

Inputs

$\log(1 + p_{\text{T}})$, η, $\sin\phi$, $\cos\phi$, $\log(1 + m)$, c_i

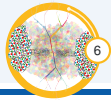
$|M| = \lceil 0.15N \rceil$ of the N valid event particles, sampled within the active p_{T} third. If that third is empty, sample from all valid particles. Final stage: all particles eligible.

Caveat

Derived mass is clipped at zero after negative raw mass-squared values; it remains a diagnostic target.

Particle directions (η, ϕ) remain visible; they are not reconstruction targets.

Full-data runs: lower reported total loss, higher momentum error



Causal Qwen

1M events: 0.761

44.1M events: 0.702

7.7% lower total held-out loss

Mamba-2 small

1M events: 0.875

44.1M events: 0.820

6.3% lower total held-out loss

PF-category loss accounts for most of the reported total-loss difference.

The $\log(1 + p_{\mathrm{T}})$ MSE increases for both models. Different test samples and training exposure prevent a controlled scaling claim.

Runs and diagnostics

← Backup index

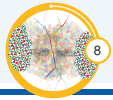
Completed Runs: Small Models

Model	Train events	Params [M]	Epochs	GPUs	Epoch [s]	Val loss	Test loss
Causal Qwen full	44.1M	1.309	115	4	2157.6	0.7023	0.7024
ParT	1M	1.307	124	4	313.4	0.7431	0.7484
Equi Qwen 1M	1M	1.308	143	4	18.8	0.7474	0.7516
Causal Qwen 1M	1M	1.309	151	4	48.7	0.7569	0.7609
ParT exact	1M	1.307	144	4	527.0	0.7593	0.7646
Mamba S full	44.1M	1.308	109	4	689.0	0.8195	0.8196
Omni S pp-HI	1M	1.573	128	4	286.5	0.8299	0.8339
Omni S HI repeat	1M	1.573	155	4	286.7	0.8363	0.8399
Omni S HI	1M	1.573	161	4	286.5	0.8383	0.8419
Mamba S 1M	1M	1.308	120	4	16.6	0.8720	0.8750
MIParT	1M	1.932	116	4	815.2	0.8771	0.8787

Timing: median final-stage epoch after the first, where available. Full-data rows cover 44.1M events, not 1M.

All recorded curricula remain in the inventory; the extra small-Omni run is not an independent initialization category.

Completed Runs: Larger Models



Model	Train events	Params [M]	Epochs	GPUs	Epoch [s]	Val loss	Test loss
Omni M pp-HI	1M	32.222	110	4	1042.3	0.8375	0.8417
Concerto	1M	207.726	128	4	532.2	0.8807	0.8841
ModernGBERT	1M	998.525	10	4	1428.2	0.8871	0.8896
Omni M HI	1M	32.222	114	4	1037.5	0.8884	0.8916

ModernGBERT completed only one epoch in each of the ten stages. Its result is a short-schedule measurement, not a converged billion-parameter baseline.

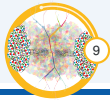
Large Mamba (1.241B) completed engineering benchmarks, not the ten-stage curriculum. Those measurements have a separate table.

Omni HI: random $\rightarrow$ PbPb; Omni pp-HI: public pp $\rightarrow$ PbPb. Pretraining provenance, not downstream mode.

Concerto; ModernGBERT; OmniLearned.

Model size, data exposure and convergence must be read together.

Completed Curricula: Held-Out Loss Components (1/2)



Model	Val total	Test total	Test log-p_{T} MSE	Test log-m MSE	Test PID CE
Causal Qwen full	0.7023	0.7024	0.06862	0.00353	0.66634
ParT	0.7431	0.7484	0.05770	0.00351	0.71778
Equi Qwen 1M	0.7474	0.7516	0.05856	0.00367	0.72043
Causal Qwen 1M	0.7569	0.7609	0.06036	0.00366	0.72890
ParT exact	0.7593	0.7646	0.06073	0.00358	0.73240
Mamba S full	0.8195	0.8196	0.06521	0.00372	0.78516
Omni S pp-HI	0.8299	0.8339	0.06183	0.00363	0.80114
Omni S HI repeat	0.8363	0.8399	0.06062	0.00363	0.80779

$$\mathcal{L}_{\text{stage }10} = \frac{1}{2}\left(\mathrm{MSE}_{\log(1+p_{\mathrm{T}})} + \mathrm{MSE}_{\log(1+m)}\right) + \mathrm{CE}_{\mathrm{PF\ PID}}$$

Sorted by final-stage validation loss. Test is evaluated at the validation-selected checkpoint.

Full-data rows use a different test view; this table is an inventory, not a common-test architecture leaderboard.

Completed Curricula: Held-Out Loss Components (2/2)

Model	Val total	Test total	Test log-p_{T} MSE	Test log-m MSE	Test PID CE
Omni M pp-HI	0.8375	0.8417	0.06326	0.00366	0.80819
Omni S HI	0.8383	0.8419	0.05869	0.00362	0.81077
Mamba S 1M	0.8720	0.8750	0.05970	0.00372	0.84332
MIParT	0.8771	0.8787	0.06489	0.00370	0.84442
Concerto	0.8807	0.8841	0.06262	0.00390	0.85086
ModernGBERT	0.8871	0.8896	0.06610	0.00381	0.85468
Omni M HI	0.8884	0.8916	0.06966	0.00386	0.85482

$$\mathcal{L}_{\text{stage }10} = \tfrac{1}{2}\left(\mathrm{MSE}_{\log(1+p_{\mathrm{T}})} + \mathrm{MSE}_{\log(1+m)}\right) + \mathrm{CE}_{\mathrm{PF\ PID}}$$

Together the two tables cover all 15 retained ten-stage reports, including Equi Qwen, the repeated Omni run and short ModernGBERT.

The 1.241B Mamba benchmarks have no validation/test result and are not assigned one here.

Reconstruction metrics: common events, different sampled masks

Model	Curriculum test loss	$\log(1 + p_{\mathrm{T}})$ RMSE	$\log(1 + m)$ RMSE	PID accuracy
Causal Qwen	0.7609	0.2461	0.0605	53.9%
Compiled ParT	**0.7484**	**0.2406**	**0.0593**	**58.9%**
Mamba-2 small	0.8750	0.2444	0.0610	54.1%

Same 250k events and 15% masking rule; sampled masks differ for Mamba.

7.58M masks per model; four layers, approximately 1.31M parameters, 1M training events.

Total loss: original curriculum test. RMSE and PID accuracy: separate saved reconstruction evaluation.

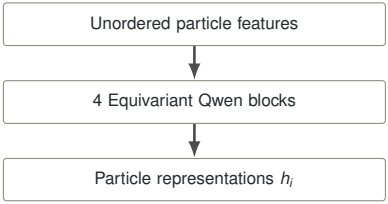
Original Model: Equivariant Qwen

Keep the optimized building blocks

RMSNorm, SwiGLU, residual connections, grouped-query attention and fused projections.

Remove sequence order

No rotary or absolute positional embeddings. No causal mask. Full bidirectional, packed-event attention.



$$F(PX) = PF(X)$$

Symmetric pooling gives an invariant event representation. Compiled BF16 / FlashAttention-3 validated; strict FP32 and baseline-relative BF16 permutation checks passed.

1,307,792 parameters; 1M training events; 143 epochs. Final-stage validation 0.7474, test 0.7516; 18.8 s/train epoch on 4 H200s.

Completed curriculum and all four independent downstream fine-tunings, with scratch controls.

Original Model: Causal Qwen

Retained from Qwen3.5

Three Gated DeltaNet blocks, then one full-attention block. Width 128; eight query heads and four key/value heads in full attention.

RMS normalization, gated feed-forward layers, residual paths, causal processing and partial rotary position encoding.

Qwen3.5 architecture documentation.

Changed for particle learning

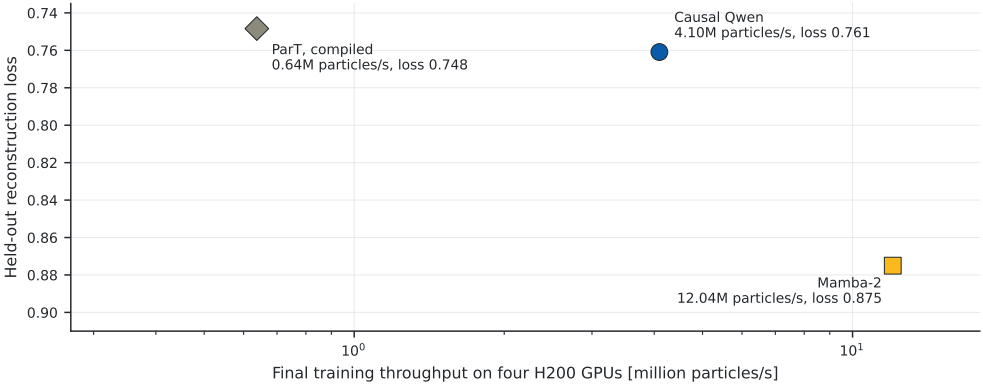
Text lookup bypassed by continuous-feature and PF-category embeddings. Language output removed; per-particle kinematic and category heads attached.

Random initialization, no pretrained language weights. Ten-stage PbPb curriculum on the 1M or 44.1M training split.

1,308,501 parameters. BF16, compiled training, optimized recurrent kernels; recorded full attention backend: PyTorch SDPA.

The contribution is the particle-level adaptation and native ion training, not a new Qwen backbone.

Reconstruction Quality and Speed: Three Small Models



ParT

lowest loss in this three-model comparison

Causal Qwen

intermediate runtime and reconstruction loss

Mamba-2

lower runtime, higher reconstruction loss

Transfer diagnostics

← Backup index

Downstream Tasks: Initialization Routes

Display label	Weight path	Campaign status
Scratch	random → supervised downstream	Completed results available
pp / FT	public pp → downstream fine-tuning	Completed results available
HI / FT	random → PbPb → downstream fine-tuning	Completed results available
pp-HI / FT	public pp → PbPb → downstream fine-tuning	Completed results available
HI / frozen	PbPb encoder fixed → trained task heads	Joint diagnostics completed
pp-HI / frozen	pp-adapted PbPb encoder fixed → task heads	Joint diagnostics completed

Label format: **model size + pretraining route** / **downstream mode**. FT means full-encoder fine-tuning.

For example, **Omni S pp-HI** / **frozen** is not a direct pp transfer or a downstream fine-tuning result.

Causal Qwen: trained order-dependent hybrid. **Equivariant Qwen (Equi Qwen)**: position-free, bidirectional; 1M curriculum and four FT/scratch tasks completed.

Two separate axes: where the encoder learned, and whether it adapts to the downstream task.

OmniLearned: What Has Actually Been Trained?

Four completed PbPb curricula, plus one retained Small HI repeat

Label	Encoder initialization and pretraining	Small test loss	Medium test loss
HI	random weights → PbPb curriculum	0.8419	0.8916
pp-HI	public pp weights → PbPb curriculum	0.8339	0.8417

Completed downstream diagnostics

Small HI / frozen and Small pp-HI / frozen. The encoder is fixed; all three downstream heads are trained jointly.

The additional Small HI run is a repeat of the native-PbPb route, not another initialization strategy.

Separate-task results available

Small: HI / FT, direct pp / FT and pp-HI / FT completed all four tasks.

Medium: HI / FT, direct pp / FT and pp-HI / FT now complete on all four tasks.

All four scratch controls are complete for both sizes.

OmniLearned; local reports verify which runs loaded public weights. All four primary curricula use the 1M training split.

Completed end-to-end comparisons are shown separately from the frozen joint-task diagnostics.

Downstream Results: OmniLearned Initialization Routes

Model / PT route	PT test loss	PID [%] ↑	Jet AUC ↑	p_{T} log-MAE ↓	JEWEL AUC ↑
Omni S, PbPb 1M	0.8399	88.59 / 88.60	0.9594 / 0.9596	0.2689 / 0.2674	0.7509 / 0.7267
Omni S, pp	—	88.61 / 88.60	0.9595 / 0.9596	0.2693 / 0.2674	0.7480 / 0.7267
Omni S, pp to PbPb	0.8339	88.60 / 88.60	0.9596 / 0.9596	0.2681 / 0.2674	0.7593 / 0.7267
Omni M, PbPb 1M	0.8916	88.51 / 88.59	0.8704 / 0.9596	0.3944 / 0.2684	0.6911 / 0.7659
Omni M, pp	—	88.61 / 88.59	0.9595 / 0.9596	0.2688 / 0.2684	0.6923 / 0.7659
Omni M, pp to PbPb	0.8417	88.43 / 88.59	0.9544 / 0.9596	0.2974 / 0.2684	0.7359 / 0.7659

Task cells: **FT** / **Scratch**. Fine-tuning updates encoder and task head; scratch starts from random weights.

Each architecture shares one scratch control per task across its pretraining routes.

Small PbPb uses the repeat-curriculum checkpoint 143335, not the earlier native-PbPb run.

For Omni Medium, direct pp leads its CMS FT routes; pp-to-PbPb improves JEWEL but trails scratch.

Downstream Results: Separate-Task Test Losses

One trained model per task. Cells: **FT** / **Scratch**; do not sum columns. CE: cross-entropy; BCE: binary cross-entropy.

Model / PT route	PID CE	Jet BCE	p_T log-MSE	JEWEL weighted BCE
Causal Qwen, PbPb 1M	0.2886 / 0.2884	0.2296 / 0.2254	0.1685 / 0.1676	0.5403 / 0.5386
Causal Qwen, PbPb full	0.2891 / 0.2884	0.2304 / 0.2254	0.1681 / 0.1676	0.5303 / 0.5386
Equi Qwen, PbPb 1M	0.2847 / 0.2846	0.2235 / 0.2218	0.1654 / 0.1637	0.5322 / 0.5367
Mamba S, PbPb 1M	0.2871 / 0.2869	0.2313 / 0.2326	0.1713 / 0.1717	0.5766 / 0.5774
Mamba S, PbPb full	0.2865 / 0.2869	0.2326 / 0.2326	0.1733 / 0.1717	0.5589 / 0.5774
ParT, PbPb 1M	0.2830 / 0.2831	0.2171 / 0.2186	0.1651 / 0.1672	0.5101 / 0.5074
MIParT, public to PbPb	— / 0.2893	— / 0.2412	0.2220 / 0.2285	0.4987 / 0.6622
Omni S, PbPb 1M	0.2829 / 0.2831	0.2138 / 0.2134	0.1581 / 0.1566	0.5227 / 0.5579
Omni S, pp	0.2826 / 0.2831	0.2135 / 0.2134	0.1581 / 0.1566	0.5144 / 0.5579
Omni S, pp to PbPb	0.2828 / 0.2831	0.2135 / 0.2134	0.1581 / 0.1566	0.5156 / 0.5579
Omni M, PbPb 1M	0.2865 / 0.2833	0.3333 / 0.2133	0.2788 / 0.1576	0.6304 / 0.5314
Omni M, pp	0.2828 / 0.2833	0.2137 / 0.2133	0.1576 / 0.1576	0.5995 / 0.5314
Omni M, pp to PbPb	0.2891 / 0.2833	0.2264 / 0.2133	0.1860 / 0.1576	0.5398 / 0.5314
Concerto, PbPb 1M	0.3319 / 0.3021	0.6291 / 0.5913	0.3627 / 0.3696	— / —

Test uses the validation-selected checkpoint. Momentum trains with log-MSE; the headline error is log-MAE.

JEWEL uses native jet weights and a separate test cohort. — means no completed evaluation.

Downstream Results: Full Four-Task Table

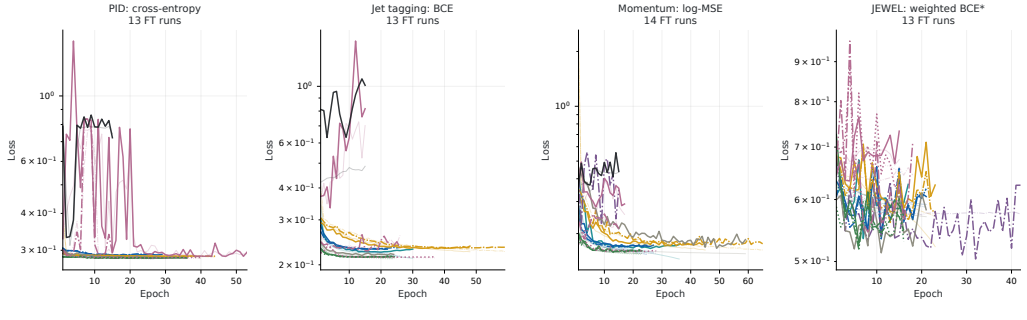
Model / PT route	PT test loss	PID [%] ↑	Jet AUC ↑	p_T log-MAE ↓	JEWEL AUC ↑
Causal Qwen, PbPb 1M	0.7609	88.46 / 88.45	0.9536 / 0.9554	0.2791 / 0.2810	0.7357 / 0.7321
Causal Qwen, PbPb full	0.7024	88.43 / 88.45	0.9534 / 0.9554	0.2825 / 0.2810	0.7325 / 0.7321
Equi Qwen, PbPb 1M	0.7516	88.56 / 88.56	0.9557 / 0.9565	0.2768 / 0.2759	0.7349 / 0.7362
Mamba S, PbPb 1M	0.8750	88.52 / 88.53	0.9528 / 0.9524	0.2841 / 0.2876	0.7232 / 0.7282
Mamba S, PbPb full	0.8196	88.54 / 88.53	0.9522 / 0.9524	0.2881 / 0.2876	0.7278 / 0.7282
ParT, PbPb 1M	0.7484	88.58 / 88.59	0.9585 / 0.9584	0.2805 / 0.2847	0.7600 / 0.7538
MIParT, public to PbPb	0.8787	— / 88.48	— / 0.9551	0.3498 / 0.3575	0.7732 / 0.4996
Omni S, PbPb 1M	0.8399	88.59 / 88.60	0.9594 / 0.9596	0.2689 / 0.2674	0.7509 / 0.7267
Omni S, pp	—	88.61 / 88.60	0.9595 / 0.9596	0.2693 / 0.2674	0.7480 / 0.7267
Omni S, pp to PbPb	0.8339	88.60 / 88.60	0.9596 / 0.9596	0.2681 / 0.2674	0.7593 / 0.7267
Omni M, PbPb 1M	0.8916	88.51 / 88.59	0.8704 / 0.9596	0.3944 / 0.2684	0.6911 / 0.7659
Omni M, pp	—	88.61 / 88.59	0.9595 / 0.9596	0.2688 / 0.2684	0.6923 / 0.7659
Omni M, pp to PbPb	0.8417	88.43 / 88.59	0.9544 / 0.9596	0.2974 / 0.2684	0.7359 / 0.7659
Concerto, PbPb 1M	0.8841	87.71 / 88.34	0.4928 / 0.8027	0.4837 / 0.4805	— / —

FT / **Scratch**; shared controls; — incomplete. Focus: PbPb 1M rows. Other PT routes: context.

Single runs. CMS / JEWEL: separate tests; JEWEL uses native weights.

Scratch is competitive. Transfer gains depend on model and task.

Downstream Results: Fine-Tuning by Task



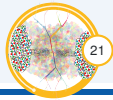
Thin / faint: training. Thick / opaque: validation. Color and dash pattern identify model / initialization route.

*JEWEL: train uses class-normalized weights; validation uses native jet weights. Loss normalizations differ.

Faint: training loss averaged over changing weights. Opaque: validation with the end-of-epoch model.

JEWEL training uses class-normalized weights; validation uses native weights. The gap is not a direct overfitting measure.

Downstream Results: Jet Tagging Full Metrics



Matched versus unmatched jets

ROC AUC ↑

Model / PT route	FT	Scratch
Causal Qwen, PbPb 1M	0.9536	0.9554
Causal Qwen, PbPb full	0.9534	0.9554
Equi Qwen, PbPb 1M	0.9557	0.9565
Mamba S, PbPb 1M	0.9528	0.9524
Mamba S, PbPb full	0.9522	0.9524
ParT, PbPb 1M	0.9585	0.9584
Omni S, PbPb 1M	0.9594	0.9596
Omni S, pp	0.9595	0.9596
Omni S, pp to PbPb	0.9596	0.9596
Omni M, PbPb 1M	0.8704	0.9596
Omni M, pp	0.9595	0.9596
Omni M, pp to PbPb	0.9544	0.9596
Concerto, PbPb 1M	0.4928	0.8027
MIParT, scratch only	—	0.9551

AUC = 0.5: chance-level.

Shared scratch; — incomplete. Single runs.

Background rejection

Background rejection at 50% signal efficiency ↑

Model / PT route	FT	Scratch
Causal Qwen, PbPb 1M	888.1	1018.2
Causal Qwen, PbPb full	910.6	1018.2
Equi Qwen, PbPb 1M	1060.5	1160.4
Mamba S, PbPb 1M	841.9	771.1
Mamba S, PbPb full	758.3	771.1
ParT, PbPb 1M	1326.7	1373.2
Omni S, PbPb 1M	1622.6	1709.6
Omni S, pp	1662.7	1709.6
Omni S, pp to PbPb	1698.3	1709.6
Omni M, PbPb 1M	253.6	1626.8
Omni M, pp	1576.2	1626.8
Omni M, pp to PbPb	739.8	1626.8
Concerto, PbPb 1M	1.1	50.4
MIParT, scratch only	—	667.4

Count-smoothed $1/\mathrm{FPR}$.

Scratch AUC: 0.80–0.96. No consistent fine-tuning gain.

Downstream Results: Jet Momentum Full Metrics

Momentum error

Log-momentum MAE ↓

Model / PT route	FT	Scratch
Causal Qwen, PbPb 1M	0.2791	0.2810
Causal Qwen, PbPb full	0.2825	0.2810
Equi Qwen, PbPb 1M	0.2768	0.2759
Mamba S, PbPb 1M	0.2841	0.2876
Mamba S, PbPb full	0.2881	0.2876
ParT, PbPb 1M	0.2805	0.2847
MIParT, public to PbPb	0.3498	0.3575
Omni S, PbPb 1M	0.2689	0.2674
Omni S, pp	0.2693	0.2674
Omni S, pp to PbPb	0.2681	0.2674
Omni M, PbPb 1M	0.3944	0.2684
Omni M, pp	0.2688	0.2684
Omni M, pp to PbPb	0.2974	0.2684
Concerto, PbPb 1M	0.4837	0.4805

Error in $\log(1 + p_{\mathrm{T}})$; clean signal jet, not pre-quenching momentum.

1M/full: pretraining split. Shared scratch; completed runs only.

Calibration

Mean response bias [%]; zero is ideal

Model / PT route	FT	Scratch
Causal Qwen, PbPb 1M	11.99	8.21
Causal Qwen, PbPb full	10.61	8.21
Equi Qwen, PbPb 1M	9.83	12.31
Mamba S, PbPb 1M	12.83	9.29
Mamba S, PbPb full	10.97	9.29
ParT, PbPb 1M	13.70	18.99
MIParT, public to PbPb	22.34	23.73
Omni S, PbPb 1M	9.21	10.12
Omni S, pp	10.64	10.12
Omni S, pp to PbPb	9.71	10.12
Omni M, PbPb 1M	35.11	11.73
Omni M, pp	10.39	11.73
Omni M, pp to PbPb	15.59	11.73
Concerto, PbPb 1M	12.19	-8.39

$100\langle p_{\mathrm{T}}^{\mathrm{pred}}/p_{\mathrm{T}}^{\mathrm{gen}} - 1\rangle$; zero is ideal.

Bias: -8.4% to $+35.1\%$. Calibration remains model-dependent.

Downstream Results: JEWEL MR–UE

Weighted ROC AUC ↑

Model / PT route	FT	Scratch
Causal Qwen, PbPb 1M	0.7357	0.7321
Causal Qwen, PbPb full	0.7325	0.7321
Equi Qwen, PbPb 1M	0.7349	0.7362
Mamba S, PbPb 1M	0.7232	0.7282
Mamba S, PbPb full	0.7278	0.7282
ParT, PbPb 1M	0.7600	0.7538
MIParT, public to PbPb	0.7732	0.4996
Omni S, PbPb 1M	0.7509	0.7267
Omni S, pp	0.7480	0.7267
Omni S, pp to PbPb	0.7593	0.7267
Omni M, PbPb 1M	0.6911	0.7659
Omni M, pp	0.6923	0.7659
Omni M, pp to PbPb	0.7359	0.7659

Weighted binary cross-entropy ↓

Model / PT route	FT	Scratch
Causal Qwen, PbPb 1M	0.5403	0.5386
Causal Qwen, PbPb full	0.5303	0.5386
Equi Qwen, PbPb 1M	0.5322	0.5367
Mamba S, PbPb 1M	0.5766	0.5774
Mamba S, PbPb full	0.5589	0.5774
ParT, PbPb 1M	0.5101	0.5074
MIParT, public to PbPb	0.4987	0.6622
Omni S, PbPb 1M	0.5227	0.5579
Omni S, pp	0.5144	0.5579
Omni S, pp to PbPb	0.5156	0.5579
Omni M, PbPb 1M	0.6304	0.5314
Omni M, pp	0.5995	0.5314
Omni M, pp to PbPb	0.5398	0.5314

Test: 633,733 jets; native jet weights, local grouped split. Split and protocol differ from the paper.
FT updates encoder + task head. Shared scratch references; — incomplete. Single runs, not seed averages.

1M/full: pretraining split. MIParT scratch is near chance (AUC 0.4996); its large FT gap needs a stronger control and repeat runs.

Omni Small improves over scratch on JEWEL; Omni Medium does not in these runs.

Downstream Results: Particle Identity Full Metrics

Overall accuracy

Accuracy [%] ↑

Model / PT route	FT	Scratch
Causal Qwen, PbPb 1M	88.46	88.45
Causal Qwen, PbPb full	88.43	88.45
Equi Qwen, PbPb 1M	88.56	88.56
Mamba S, PbPb 1M	88.52	88.53
Mamba S, PbPb full	88.54	88.53
ParT, PbPb 1M	88.58	88.59
Omni S, PbPb 1M	88.59	88.60
Omni S, pp	88.61	88.60
Omni S, pp to PbPb	88.60	88.60
Omni M, PbPb 1M	88.51	88.59
Omni M, pp	88.61	88.59
Omni M, pp to PbPb	88.43	88.59
Concerto, PbPb 1M	87.71	88.34
MIParT, scratch only	—	88.48

Generator labels; PF category remains an input.

1M/full: pretraining split. Shared scratch; — incomplete. Single runs.

Balance across classes

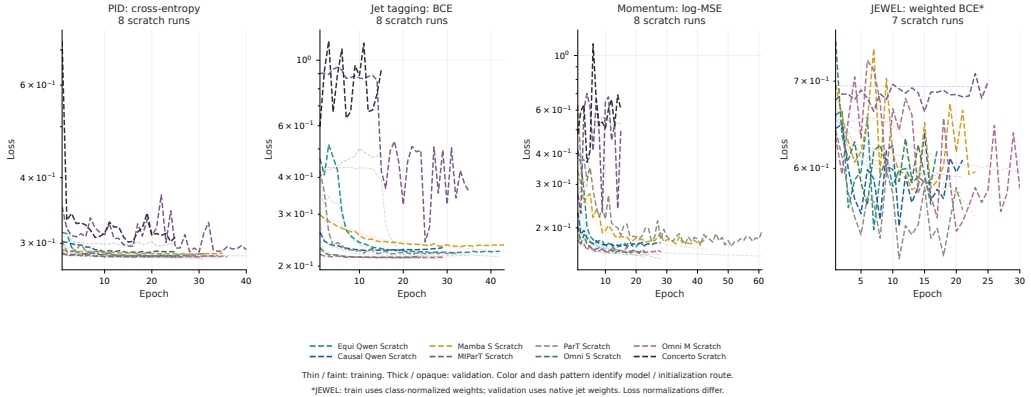
Mean five-class efficiency [%] ↑

Model / PT route	FT	Scratch
Causal Qwen, PbPb 1M	54.17	52.22
Causal Qwen, PbPb full	54.61	52.22
Equi Qwen, PbPb 1M	55.90	54.91
Mamba S, PbPb 1M	55.46	54.96
Mamba S, PbPb full	55.08	54.96
ParT, PbPb 1M	55.78	55.85
Omni S, PbPb 1M	54.06	54.61
Omni S, pp	55.67	54.61
Omni S, pp to PbPb	55.47	54.61
Omni M, PbPb 1M	54.12	54.69
Omni M, pp	56.88	54.69
Omni M, pp to PbPb	44.86	54.69
Concerto, PbPb 1M	46.07	53.91
MIParT, scratch only	—	54.35

Unweighted mean class recall.

Overall accuracy hides weak electron and muon recall.

Downstream Results: Scratch Learning Curves by Task



Thin / faint: training. Thick / opaque: validation. Color and dash pattern identify model / initialization route.

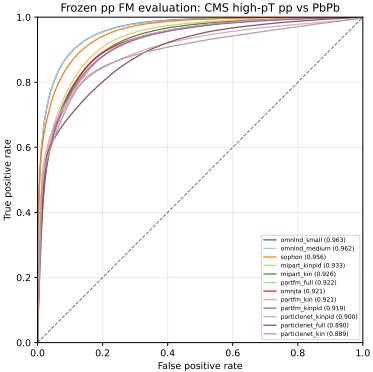
*JEWEL: train uses class-normalized weights; validation uses native jet weights. Loss normalizations differ.

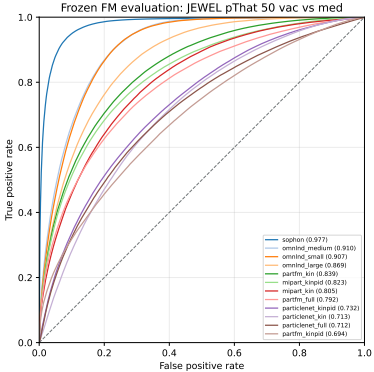
All completed scratch controls; encoder and task head train from random weights. Faint: training; opaque: validation.

Full recorded histories shown, including runs beyond 20 epochs. JEWEL train/validation weight normalizations differ.

A completed run need not learn useful discrimination: MIParT JEWEL scratch remains near chance.

Historical Frozen-Encoder Classification: CMS and JEWEL

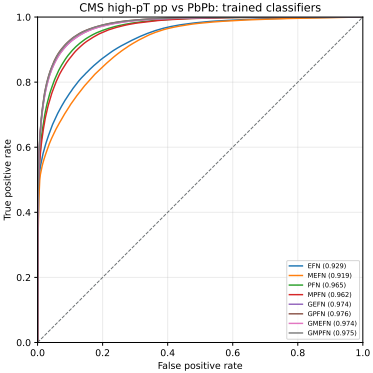


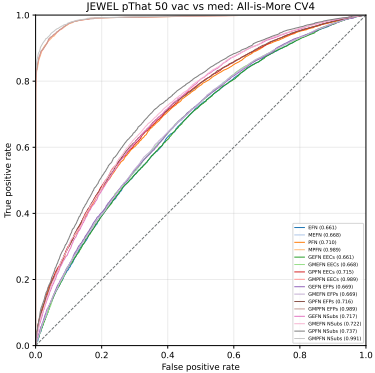


Trained readouts on fixed pp encoders distinguish CMS pp/PbPb data and JEWEL vacuum/medium jets.

June jet-level studies, separate selections and splits. Discrimination does not isolate medium effects: backgrounds and kinematics can contribute. OmniLearned.

Historical EFN-Family Classifiers: May–June 2026





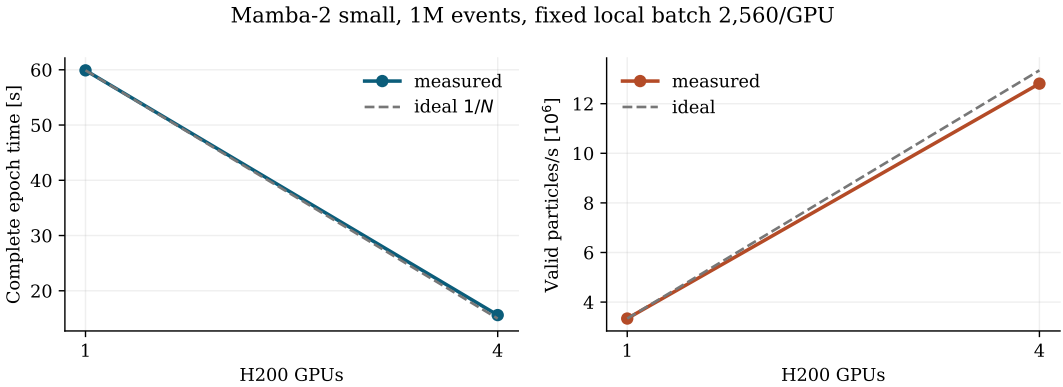
Earlier jet-level classifiers: CMS high-p_{T} pp versus PbPb data; JEWEL vacuum versus medium-modified jets.

Different selections and split protocols: do not compare these AUCs directly with the current benchmark. Energy Flow Networks.

Compute and scale

← Backup index

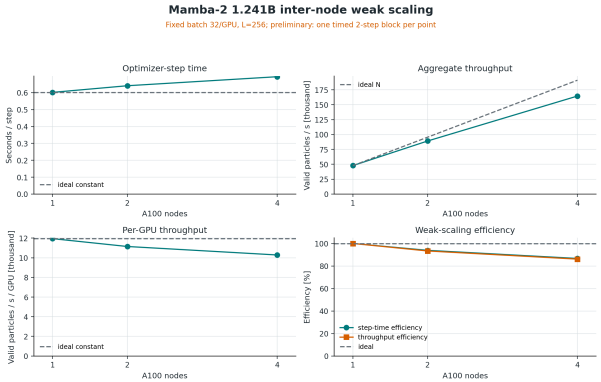
Measured H200 Intra-Node Scaling



H200, fixed local batch 2,560/GPU. One to four GPUs on one node.

Existing exploratory measurements; data-order configurations differ between some runs.

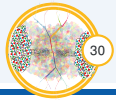
Preliminary A100 Inter-Node Scaling



A100, fixed local batch 32/GPU; global batch grows with node count.

Short timing windows, not full epochs; remeasure performance on any Tier-0 or Tier-1 cluster.

Large Mamba: Engineering Benchmarks, Not Full Curricula



Official Mamba-2 1.3B backbone, 1.241B parameters after particle adaptation

48 layers, width 2,048. All runs below: 1M training events, one H200.

Benchmark	Job	Epochs	Local batch	Epoch [s]	Particles/s
Initial full epochs	120722	3	32	10765.0	18,563
Optimized full epoch	131996	1	288	7983.8	25,029
Multiplicity-ordered epoch	131998	1	288	8109.4	24,642
Two-epoch MFU benchmark	143030	2	288	7961.5	25,099

The job labelled curriculum (131998) requested one epoch ordered by event multiplicity. It completed normally; no checkpoint was saved.

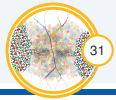
That ordering is not the ten-stage p_T/PID curriculum. No completed large-Mamba ten-stage training or validation/test result was found.

Epoch time excludes the first epoch when possible; single-epoch times include its overhead. The objectives differ from the mass curriculum.

Mamba-2 paper; Official 1.3B configuration.

A 1.24B Mamba model ran 1M-event epochs on one H200; ten-stage training and downstream tests remain.

Benchmark: One Protocol Across Collision Systems



Start with public samples; extend only where the inputs support the task.

1. Available samples

CMS: pp, pPb, PbPb
Collision data + simulations

ATLAS PbPb: tracks + MC

ALICE: 2015 PbPb

LHCb pp: custom ntuples

Different formats and acceptances; validate each task's inputs.

CMS releases; ATLAS PbPb; ALICE PbPb; LHCb pp.

2. Common tests

Tracks, activity and flow

Define shared observables

Fix event splits and selections

Correct detector effects

Scratch versus pretraining

Match labels and total compute

Repeat seeds; test MC closure

Proposed protocol; task-specific baselines.
Flow reference methods.

3. Access gates

Light ions: O–O / Ne–Ne
Event-level public release not verified.

Jets need suitable collections
ATLAS HION14: no full jet or PF collections. Use minimum-bias tracks.

Unfolding needs closure
Released observable tables are not this encoder's particle inputs.

Light-ion run; ATLAS formats; OmniFold-HI data.

A shared protocol, not one pooled dataset.